The environment in Tanzania as a source of *Mycobacterium tuberculosis* complex and diversity analysis of the slow growing mycobacteria.

by

Yu-Jiun Hung

A thesis submitted to the University of Warwick for the degree of **Doctor of Philosophy**

School of Life Sciences

University of Warwick,

Coventry, CV4 7AL,

United Kingdom

July 2015

Contents

Lists of Figures

List of Tables

Acknowledgements

This project was financially supported by NIH program grant. I would like to express my sincere thanks to Professor Elizabeth Wellington and Dr. Orin Courtenay for providing me with all the necessary guidance, support and facilities for this study during my time at University of Warwick.

I appreciate past and present research team members of Professor Elizabeth Wellington's group for their guidance and advice for my research, in particular Dr. David Porter, Dr. Emma Travis and Dr. Philip James. I also wish to thank Hayley King and Andrew Murphy for constantly helping me in my research and proof reading my thesis. I acknowledge the team members at Sokoine University of Agriculture in Tanzania, particularly Goodluck Paul and Joseph Malakalinga for their guidance and assistance during sample collection.

Finally I would like to take this opportunity to express gratitude to beloved partner Clare Tu for her continuous support and to my parents, Jheng-Jhih Hung and Bo-He He for constant financial support and encouragement.

Declaration

I here declare all my results observed in this these was presented by myself under the supervision of Professor Elizabeth Wellington and Dr. Orin Courtenay, unless stated otherwise. This thesis has not been summited for any other degree in other academic institutions. All source of information presented in this thesis have been acknowledged by the reference.

Summary

Tuberculosis (TB) is the most prevalent infectious human disease and health burden worldwide. There are over eight million incident cases of TB and this has caused more than one million deaths globally. In addition, bovine Tuberculosis (bTB) has become widespread with the disease endemic in most African countries, including Tanzania. For this reason, detection and treatment for the two causal agents *Mycobacterium tuberculosis* (Mtb) and *Mycobacterium bovis* (Mb) has become a priority in Tanzania. Other slow growing mycobacteria (SGM) are also responsible for respiratory disease in humans and animals. This project focuses on the environment as a potential reservoir of Mtb, Mb and other SGM by comparison of samples from different sites and use of case controls which can help to establish if shedding correlates with disease and indeed can be a potential source of continuing infection. In this study, IMC-qPCR and different specific primers RD4 scar and RD9 were developed to detect and quantify Mtb and Mb in a range of environmental samples taken during the dry and wet season from villages in Tanzania. Mtb and Mb were both detected in cattle faecal samples taken from animals kept by the same household. For the SGM prevalence and diversity specific primers were used for oligotyping following deep sequence analysis by pyrosequencing of amplicons. The oligotyping result matched the identification of bacterial species in cattle lesion and wildlife tissue. The higher prevalence of Mtb was detected from households with TB patients compared to control households. This work is part of a collaboration with UCD and USF in USA and SUA in Tanzania so data on infection status of human and animal populations will be made available following approvals then compared with environmental reservoirs.

List of Abbreviations

ANOSIM	Analysis of variance of similarity
ANOVA	Analysis of variance
AIDS	acquired immune deficiency syndrome
BCG	Bacillus Calmette-Guerin
BLAST	Basic Local Alignment Search Tool
bp	Base pair
bTB	Bovine Tuberculosis
dNTP	Deoxyribonucleotide triphosphate
ELISA	Enzyme-linked immunosorbent assay
EM	Environmental mycobacteria
HIV	Human immunodeficiency virus
IFN-y	Interferon Gamma
Ig	Immunoglobulin
IMC	Immunomagnetic capture
LSPN	List of Prokaryotic names with Standing in Nomenclature
MAbs	Monoclonal antibodies
MAC	*Mycobacterium avium* complex
MAP	*Mycobacterium avium subspecies paratuberculosis*
Mb	*Mycobacterium bovis*
MDR-TB	Multiple drug resistance Tuberculosis
MED	Minimum Entropy Decomposition
MG-RAST	Metagenomics RAST
MLST	Multilocus sequence typing
Mtb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
NMDS	Nonmetric Multidimensional Scaling
NIH	National Institutes of Health

NTM	Non-tuberculous mycobacteria
OTU	Operational Taxonomic Unit
PAbs	Polyclonal antibodies
PCA	Principle component analysis
PCoA	Principal coordinates analysis
PPD	Purified protein derivative
PPi	Pyrophosphate
QFT	QuantiFERON
QIIME	Quantitative Insights into Microbial Ecology
qPCR	Quantitative polymerase chain reaction
RD	Region of Difference
RDP	Ribosomal Database Project
RFLP	Restriction fragment length polymorphism
RGM	Rapid-growing mycobacteria
SGM	Slow-growing mycobacteria
TB	Tuberculosis
WHO	World Health Organisation

Chapter 1 General Introduction

1.1. Taxonomy and physiology of the *Mycobacterium* genus

The *Mycobacterium* genus belongs to a group of slow growing Gram positive bacteria family Mycobacteriaceae, order Actinomycetales, phylum Actinobacteria and kingdom Bacteria. There are 165 known *Mycobacterium* species in existence, based on the List of Prokaryotic names with Standing in Nomenclature (LSPN). Mycobacterial cells are aerobic, acid-alcohol-fast, have a G-C rich genome, and are non-motile and rod-shaped (Falkinham 2009). A cell is 1-10 µm long and 0.2-0.6 µm width. The cell wall is made up of peptidoglycan layer plus a hydrophobic thick mycolic acid layer interspersed with lipoarabinomannams and anchored by arabinogalactan polymers attached to lipomannam portion of more lipoarabinomannams attached to plasma membrane (Barry, Lee et al. 1998, Park and Bendelac 2000).

The *Mycobacterium* genus is comprised of fast and slow-growing species; fast growers are generally saprophytes found in soil and waters often associated with commercially important degradative capabilities (Howard and Byrd 2000). The slow-growing mycobacteria (SGM) include many pathogenic species such as the *Mycobacterium tuberculosis* complex (MTBC), and certain environmental mycobacteria (EM) known as non-tuberculous mycobacteria (NTM) and atypical mycobacteria are both capable of causing disease. The two best known pathogens are *M. tuberculosis* (Mtb) and *M. bovis* (Mb). However, some EM not only cause pulmonary disease with similar to tuberculosis in humans and other animals, but also skin disease and disseminated disease in humans as well as in animals.

Rapid growing mycobacteria (RGM) such as *Mycobacterium smegmatis* and *Mycobacterium fortuitum* have a doubling time of 60-90 min on plates and in media, however, SGM like MTBC members have a doubling time of approximately 24 hours (Greendyke, Rajagopalan et al. 2002). As such faster mycobacterial cells appear on selective plates within 2-7 days, while slower sub-species take 2.4 weeks (Shinnick and Good 1994). Therefore isolation of SGM bacteria is problematic and time-consuming because of growing-time, and few selective media, such as Middlebrook, exist to recover SGM cells (Young, Gormley et al. 2005). Differentiation between SGM members is also challenging. Solution to the issues in cultivation based methods are now widely discussed, one possibility is the development or reliable and rapid molecular techniques and methods for the detection of SGM species within a short time period.

Detection and identification of different *Mycobacterium* species is a critical technique for disease diagnosis and research differentiation, and the highly conserved housekeeping gene 16S rRNA is useful signature gene for general separation of each *Mycobacterium* species and had been used to effectively construct a phylogenetic tree (Figure 1.1). Nevertheless, the drawback of this 16S rRNA separation is the lack of complete differentiation of species within the SGM, namely within the MTBC and *Mycobacterium avium* complex (MAC) (Tortoli 2012). It is therefore important to develop phylogenetic methods based on other unique genes to complement 16S rRNA phylogeny to segregate each bacteria from its complex. Other genes, *rpoB* and *hsp65*, are employed for taxonomic analysis instead of 16S rRNA based methods to segregate each complex member like Mtb and Mb from MTBC (Tortoli 2012).

Moreover, *rpoB* is a gene encoding the β subunit of bacterial RNA polymerase, and can successfully discriminate Mtb from other *Mycobacterium* species using PCR mutagenesis technique (Floss and Yu 2005). Compared to *rpoB*, the *hsp65* gene encodes a 60kDa heat-shock protein, and has been recognised as an effective phylogenetic marker in identification of cultured *Mycobacterium* spp. because of its highly conserved primary structure (Kim, Kim et al. 2005).

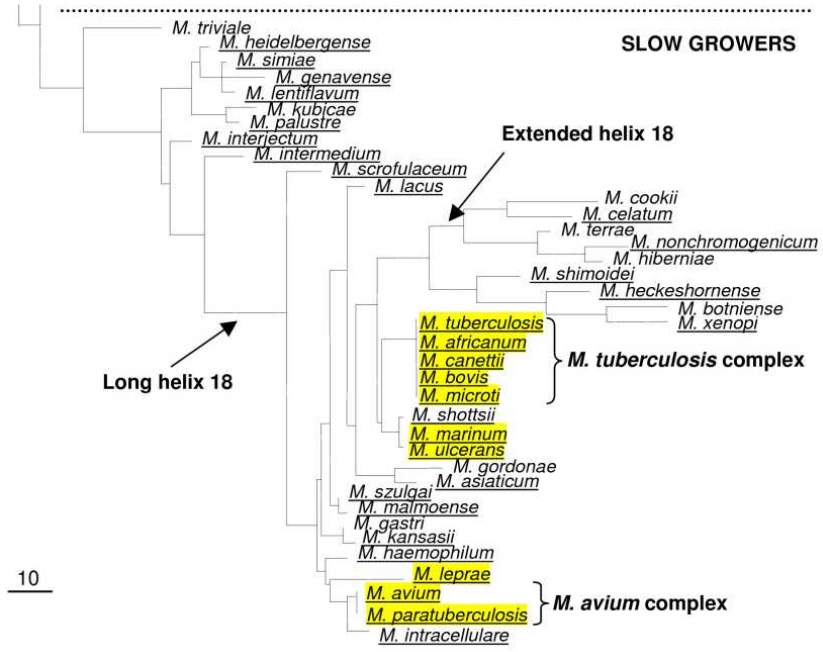


Figure 1.1 The phylogenetic tree based on 16S rRNA gene of SGM species. The members of the MTBC and the MAC are highlighted in yellow. The divisions between long helix 18 and extended helix 18 of the 16S rRNA gene sequence are indicated (Leclerc, Haddad et al. 2000, Gey van Pittius, Sampson et al. 2006).

MTBC is a species group classified within the slow growing subgroup of the *Mycobacterium* genus, consisting of nine members, *M. canettii, M. caprae, M. microti, M. pinnipedii, M. africanum, M. mungi* and well-known human and animal pathogens, Mtb and Mb respectively and a new species *M. oryxis* identified recently (van Ingen,

Rahim et al. 2012). Development of effective differentiation methods for each sub-species is highly challenging given their close similarity.

Identification and differentiation of each MTBC individual presents a significant challenge because of the high similarity between members which approaches approximately 99.95% at the nucleotide level (Garnier, Eiglmeier et al. 2003). Whole-genome DNA microarrays have been used to establish regions of difference (RD) which provide specific target regions for the use of quantitative polymerase chain reaction (qPCR) based detection and differentiation methods to separate each MTBC member from its complex (Halse, Escuyer et al. 2011). There are 16 RDs representing regions of the MTBC genome deleted in Mb Bacillus Calmette–Guérin (BCG), but which can be present in Mtb and other MTBC members (Reddington, O'Grady et al. 2011). For instance, RD9 is present in the genome of Mtb and *M. cannettii* only (Table 1) (Brosch, Gordon et al. 2000), however, it can be used to differentiate Mtb from other MTBC members as *M. canettii* is geographically limited to Djibouti in the horn of Africa (Gutierrez, Brisse et al. 2005). There is currently a real-time qPCR test which can diagnose TB while differentiating between Mtb and *M. canettii* using a specific primer called $RD^{canettii1}$ developed by Reddington *et al* at 2011. This novel primer is designed from the flanking region RD12 which is deleted in *M. canettii* but present in Mtb (Halse, Escuyer et al. 2011).

Table 1.1 Signature patterns of positive and negative qPCR results used to determine MTBC species targeting the regions of deletion (Halse, Escuyer et al. 2011)

Organism	PCR result for target:				
	RD1	RD4	RD9	RD12	ext-RD9
M. tuberculosis	+	+	+	+	+
M. bovis	+	−	−	−	+
M. bovis BCG	−	−	−	−	+
M. africanum	+	+	−	+	+
M. microti	−	+	−	+	+
M. canettii	+	+	+	−	+
NTM[a]	−	−	−	−	−

[a] NTM, nontuberculous mycobacteria.

1.2. Human and livestock disease report associated to *Mycobacterium* species

Mtb is the most important pathogen within MTBC and contributes to almost 95% of human tuberculosis (TB) cases in both adults and children worldwide (Cosivi, Grange et al. 1998). The high prevalence and mortality of pulmonary disease TB is demonstrated in a report published by the WHO, which shows 8.7 million (range from 8.5 to 9.2 million) new cases in 2013 and mortality of 1.4 million people from TB including almost 1 million deaths in HIV-negative individuals and an additional 0.4 million deaths from HIV-associated TB in the world (Zumla, George et al. 2015). Even though the mortality rate of TB has fallen by 45% since 1990 and prevalence rate had decreased by 41% during the same period, there are still an estimated 9 million cases of people suffering from TB in 2013 (Zumla, George et al. 2015). High prevalence and mortality is a health burden worldwide and there is a critical funding gap of up to 8 billion US dollars per year for TB care and control and 2 billion US dollars per year for research development (Zumla, George et al. 2015). Human TB treatment and medical diagnosis are considered as a priority worldwide, especially in some undeveloped African countries. In addition, bovine TB (bTB) is an endemic disease in most African

countries, including Tanzania (de la Rua-Domenech, Goodchild et al. 2006). Another serious health burden with high prevalence and mortality in Africa is HIV/AIDS. This immunocompromised human population (Etter, Donado et al. 2006) makes up approximately 40% of the TB patient population associated with HIV in Africa compared to only 14% worldwide (Havlir, Getahun et al. 2008). In Tanzania, nearly 38% of TB cases also suffer from HIV infection, about 2.5 times more than the global rate as reported in a WHO report in 2013 (Zumla, George et al. 2015). Therefore the development and optimisation of molecular diagnosis and detection methods for the two causal agents, Mtb and Mb, has become a strategic policy of governments in these African countries.

High mortality and morbidity in bTB infected animals also appears to be occurring in livestock and wildlife animals such as cattle and buffalo in Africa (Shitaye JE 2007, Humblet, Boschiroli et al. 2009). Similarly, wildlife such as foxes and badgers are a potential disease reservoir for livestock which were diagnosed with bTB infection based on the skin test (Humblet, Boschiroli et al. 2009). Some human TB cases are caused by Mb due to consumption of products from livestock animals, namely unsterilised milk and meat. Additionally, there is evidence that Mb cells can be detected effectively in milk using qPCR with specific primers (Zanini, Moreira et al. 1998, Roug, Perez et al. 2014). Furthermore human pulmonary TB triggered by Mb has been diagnosed in residents living in rural communities, especially pastoralists living in very close contact with livestock animals, resulting from inhalation of dust particles and bacterial-containing aerosol emitted by bTB-infected animals (Daborn,

Grange et al. 1996). This provides evidence of transmission routes correlated to animal reservoirs and indicate environmental hot-spots of disease.

There is evidence to suggest that other SGM species can be transmitted through the environment. The *Mycobacterium avium subspecies paratuberculosis* (MAP) has been shown to survive and persist in soil and root samples, and was able to infect goats which subsequently digested these roots, as demonstrated by a specific qPCR assay (Kaevska, Lvoncik et al. 2014). Multilocus Sequence Typing (MLST) and Restriction Fragment Length Polymorphism (RFLP) analysis of M. avium strains taken from humans, their livestock, and environmental reservoirs suggests these strains are similar, if not the same (Kolb, Hillemann et al. 2014).

1.3. Molecular diagnostic and detection methods for TB and bTB in human and cattle

TB diagnosis of clinical cases still relies on symptoms and obvious clinical features including weight loss, night sweats, exhaustion, and a persistent cough with phlegm containing blood in TB endemic countries. However this method of TB diagnosis cannot provide as effective and rapid detection of TB as can be achieved using molecular techniques and modern diagnostic methods (Table 1.2.). It is challenging to recognise TB when other conditions, such as AIDS or malnutrition, are causing illness with similar symptoms. In order to avoid misdiagnosis, an effective and rapid method, QuantiFERON (QFT), relies on measuring the concentration of interferon-gamma (IFN-γ) in blood. IFN-γ is released by specific T-cells into the blood system

when exposed to TB pathogen antigens, EST-6, CFP-10 and TB7.7 (Rose, Kimaro et al. 2012). This is the principle behind many immunological assays used in diagnostics (Table 1.2). However more recently DNA-based detection methods have been used to determine presence of the pathogen and are often used in combination with drug resistance assays such as in the Xpert MTB/RIF system. These molecular methods provide rapid, sensitive and specific tests and can also be used with non-clinical samples (Williamson, Basu et al. 2012).

There are still labs using traditional detection methods in some African countries and these methods normally take a week or longer to complete examination within a two-stage process. The first stage is using sputum specimen exhaled from patients and examined with acid-fast smear stain. This is problematic because all mycobacteria are acid fast, and there is a lack of sensitivity, as > 10^4 cells ml^{-1} of sputum sample are required for a positive score (Kaul 2001). The second stage is traditional selective media for cultivation which 2-4 weeks to grow bacterial cells. In comparison with the MGIT and other similar highly sensitive cultivation systems provided in modern diagnostic laboratories, traditional culture is time-consuming and lacks sensitivity. In addition, this type of cultivation technique cannot work with faecal and environmental soil samples as the bacterial load is too great for selective antibiotics (Young, Gormley et al. 2005).

Diagnosis time can be reduced from three weeks based on culture to one or two weeks based on molecular detection with a specific DNA probe method where the target bacterial cells are lysed and DNA hybridised to a species-specific fluorescently labelled probe (Table 1.1). Molecular detection was optimised for Mtb cells with high

sensitivity, specificity and rapidity but these methods cannot differentiate Mtb and Mb. Therefore it is necessary that molecular assays for detecting the presence of Mtb and Mb cells in environmental samples, such as milk, soil, faeces and water, be developed.

Isolation of SGM cells is problematic and no adequate selective media exist to recover SGM cells from environmental, especially Mtb and Mb as target cells in our investigation (Young, Gormley et al. 2005). Under environmental conditions, Mtb and Mb cells are stressed, these species become more sensitive to decontamination methods involved in their clinical isolation (JS 2003). The high sensitivity and specificity of molecular methods instead of isolation are advantages for non-invasive sampling of faecal and milk samples rather than blood and tissue means that testing is easier and in the case of wildlife faecal samples are feasible whereas trapping is not. We have demonstrated that tracheal shedding in badgers directly correlates with faecal shedding (Hayley King 2015). In this project the efficiency of selective isolation and cultivation in addition to molecular qPCR detection with specific primer/probe sets, is optimised and sensitivity increased by use of immunomagnetic capture (IMC); concentrating and quantifying target bacterial cells from samples before qPCR or cultivation.

Table 1.2 Comparison of molecular diagnostic and detective method for Tb and bTB in human and cattle

Methods	Human	References	Cattle	Reference
Culture	Specimen includes sputum, CSF, pus and biopsied tissue. Löwenstein-Jensen, Kirchner, or Middlebrook media (7H9, 7H10 and 7H11) was employed for culture	(Kumar 2007)	Specimen includes sputum, CSF, pus and biopsied tissue. Löwenstein-Jensen, Kirchner, or Middlebrook media (7H9, 7H10 and 7H11) was employed for culture	(Kumar 2007)
PCR and qPCR	Specific and sensitive gene primer and probe enable an ideal quantitative tool for detection	(Reddington, O'Grady et al. 2011)	Specific and sensitive gene primer and probe enable an ideal quantitative tool for detection	(Reddington, O'Grady et al. 2011)
ALS Assay	The secretion of antibodies from activated B cells found in blood circulation in the short period in response to TB-antigen	(Raqib, Rahman et al. 2003)		
Tuberculin skin test	Major Tuberculin skin test for humans		Major Tuberculin skin test for cattle	
Interferon γ release assays	T lymphocytes will release interferon γ when exposed to specific antigen	(Halevy, Cohen et al. 2005)		
QuantiFERON-TB	Immune response to Mtb is measured through the concentration of interferon-gamma (IFN-γ) released in whole blood circularity system.	(Gerald H. Mazurek 2003)		
QuantiFERON-TB Gold	Whole blood test for diagnosing Mtb infection including latent tuberculosis infection and TB disease.	(Mazurek, Jereb et al. 2005)		
QuantiFERON-TB	Detection of IFN-γ induced by Mtb and an enzyme linked	(Anonymous 2009)		

Gold In-Tube	immunosorbent assay (ELISA) is used to measure the amount of IFN-γ present in each of the three tubes (Nil control, TB-Antigen, Mitogen control).			
Xpert MTB/RIF	This molecular diagnosis detects DNA sequences specific for Mtb and rifampicin resistance by PCR. It uses 3 specific primers and 5 molecular probes to target *rpoB* gene.	(Helb, Jones et al. 2010)		
STAT-PAK			An immunochromatographic technology for detection of immunoglobulin A(IgA), IgM and IgG antibodies to Mb and Mtb.	(Gowtage-Sequeira, Paterson et al. 2009)
Caudal fold test			A purified protein derivative (PPD) from Mb is injected into a fold of skin at the base of tail.	(Palmer, Waters et al. 2006)
Comparative cervical test			Two sites are injected with PPD, one from Mb and one from *M. avium* to distinguish specific immune responses to Mb from non-specific.	(Palmer, Waters et al. 2006)

Molecular methods and IMC were developed to extract Mtb and Mb cells straight from soil and other media using commercial monoclonal antibodies (MAbs) and polyclonal antibodies (PAbs), and this method has been employed to capture and concentrate Mb and Mtb cells efficiently from samples (Sweeney, Courtenay et al. 2006).

Previous research indicates that Mtb and Mb can survive in environmental reservoirs such as soil for 30 days and maintain their virulence (Young, Gormley et al. 2005, Ghodbane, Mba Medie et al. 2014). The contribution of environmental reservoirs will be considered as one part of the transmission routes of TB infection in Tanzania, and establish the non-invasive molecular analysis of DNA-based target in faeces, milk, soil and water as a marker of disease incidence.

1.4. Introduction to high throughput sequencing technologies

It is not possible to delineate the composition of the whole bacterial community in the ecosystem using qPCR methods, which only target one or two bacterial species. Only a small proportion of bacterial species from soil are culturable, and as such culture is also insufficient as a measure of community composition. In addition estimations of species richness obtained from cultivation techniques are inaccurate and insensitive as a result of this large unculturable bacterial fraction and low resolution selective media. The remedy to this predicament was to provide accurate and robust methods for meta-taxonomic analysis of environmental samples.

Next generation sequencing techniques including454 pyrosequencing and Illumina MiSeq sequencing platforms have been widely applied in genetic research. 454 pyrosequencing had been developed for almost three decades beginning in 1987. This technique is based on pyrophosphate (PPi) generation upon nucleotide incorporation – the PPi is used to oxidise luciferin, generating light. As nucleotides are sequentially added and removed from the reaction, light production indicates incorporation of a given nucleotide - greater intensity indicates multiple incorporation events of the same nucleotide. (Ronaghi, Uhlen et al. 1998). Another recent sequencing approach, Illumina MiSeq, was released in 2006 and this approach uses reversibly terminal nucleotides, each labelled with a different fluorescent probe, to identify the sequence of DNA. Similar to 454, the bases are removed prior to identification of fluorescence by a charge-coupled device, and cleavage of the dye and terminal blocking group. As such, only one nucleotide can be incorporated at time, allowing accurate identification of polynucleotide tracts.(Mardis 2008).

1.4.1. 454 Roche Pyrosequencing

Pyrosequencing was based on specific amplicons with fusion primers to ensure the correct sequence size and order. The fusion primers consisted of specific primer, adaptors and ligators were required for pyrosequencing to label the amplicons in each sample set (Figure 2.9). These fusion primers were attached using PCR approach in subsequent amplicons purification, quantitation, amplification, and sequencing steps (Figure 1.2A). These amplicons were denatured and became single strand template DNA attached to a bead 28 µm in diameter (Figure 1.2B). These single DNA templates were amplified in parallel to create millions of copies on each bead. The

beads were then loaded into each well (Figure 1.2C) and DNA sequencing was begun by the introduction of four deoxynucleotide triphosphates (dNTPs). The dNTPs were sequentially added and removed from the reaction. PPi was released when the nucleic acid polymerization reaction occurred, and a nucleotide was incorporated by DNA polymerase. The ATP sulfurylase enzyme uses released PPi to convert adenosine 5' phosphosulphate to ATP, which is used by luciferase to convert luciferin to oxyluciferin, generating visible light, which is detected with a charge-coupled device. This light production is proportional to the amount of PPi release, and this, together with the sequential addition and removal of dNTPs, is what allows the sequencing of the amplicon. (Figure 1.2D) (Ronaghi, Uhlen et al. 1998).

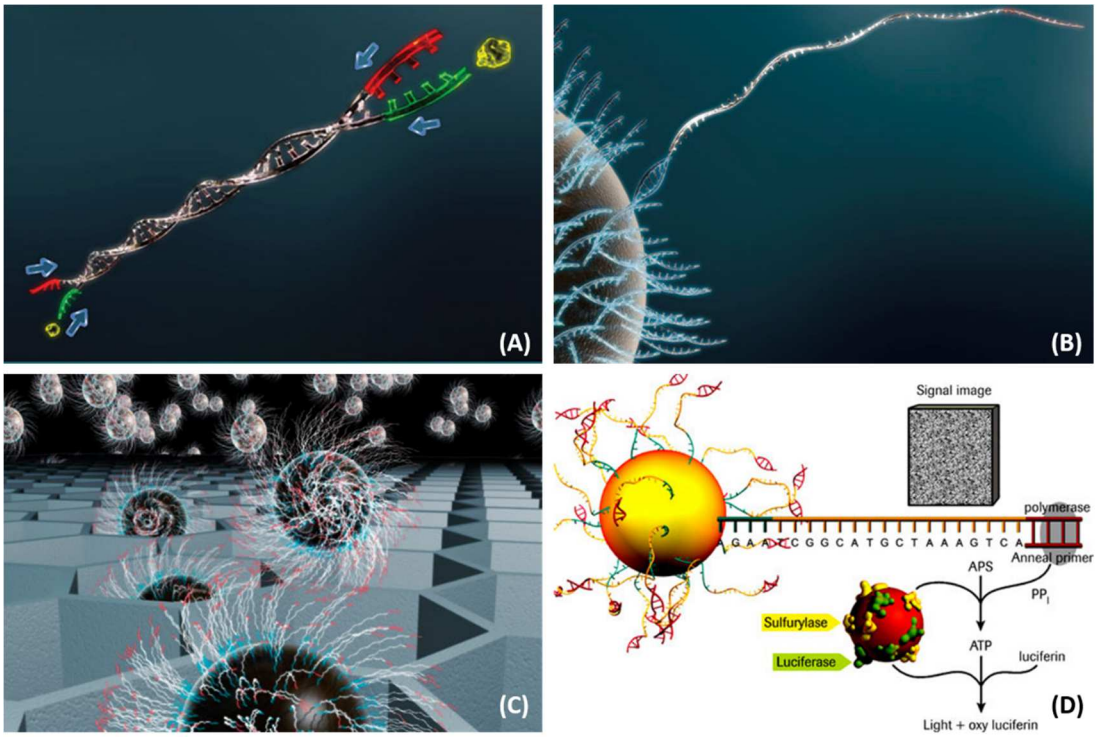


Figure 1.2 Pyrosequencing schematic of protocol taken from Roche 454 http://my454.com/products/technology.asp

Pyrosequencing has been widely applied in diverse area of research such as (1) whole genomic sequencing from sequenced fragments of DNA and genome assembly, (2) particular sequence region targeting for mutation research associated with cancer or other diseases, (3) metagenomics study using universal 16S/18S rRNA primer for microbial species identification or bio-diversity analysis of environmental study using specific sequence region primers, (4) microbial typing of mutations conferring antibiotic resistance to bacteria or differentiation of high similarity bacterial or viral strains (Marsh 2007).

The two main limitations of the pyrosequencing approach consist of the read lengths of targeting amplicons and signal to noise ratio interfering with accuracy (Hert, Fredlake et al. 2008). The GS FLX Titanium series of reagents were launched by 454 Life Sciences to enable the improvement of read lengths to 400-500 bp per read lengths with enormous 400-600 million sequences per run in 2008. In addition the signal noise produced from pyrosequencing can be improved efficiently through optimisation of the PCR process prior to pyrosequencing and rigorous validation of noise removal algorithms with test data (Quince, Lanzen et al. 2011).

1.4.2. Illumina MiSeq

The Illunina MiSeq sequencing-by-synthesis approach enabled a lower per bp error rate, similar read length (2X300 bp), higher quality and lower cost compared to 454 pyrosequencing (Nelson, Morrison et al. 2014). MISeq relies on detection of fluorescence generated during DNA amplification, using four modified dNTPs with unique fluorescence markers. These modified dNTPs prevent sequential addition of further dNTPs, as they possess cleavable terminal blockers in addition to a

fluorescent marker. As such only one dNTP residue is incorporated at a time. The MiSeq reaction thus incorporates one dNTP, removes all others and measures fluorescence, then cleaves the fluorescent marker and terminal blocker, allowing the addition of the next nucleotide. (Figure 1.3).

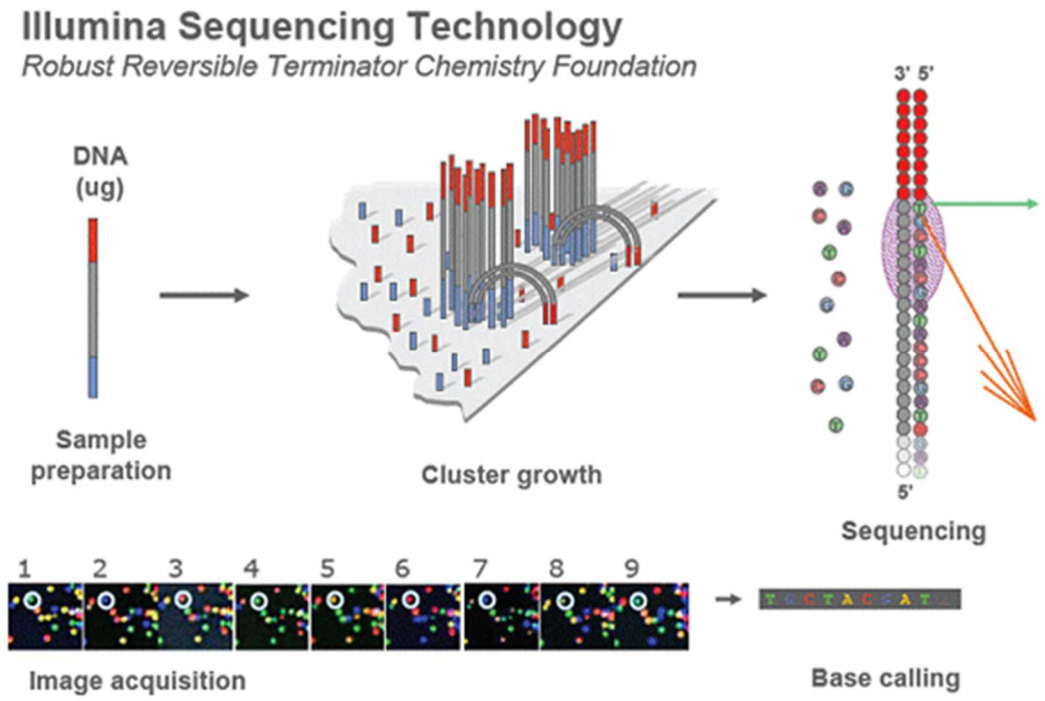


Figure 1.3 The processes of MiSeq approach for sequence synthesis and signal measure. http://www.illumina.com/technology/next-generation-sequencing.html

Applications of MiSeq sequencing are similar to pyrosequencing, both of them have been applied in mutation detection, bacterial and viral species identification in diversity analysis, microbial antibiotic resistance study, whole genomic sequencing and assembly and metagenomics research. The comparisons between these two NGS approaches have been widely discussed and are summarised in Table 1.3 (Gibson, Shokralla et al. 2014, Nelson, Morrison et al. 2014, Smith and Peay 2014). Miseq achieved high number of target reads, 2 x 300 bp read length, medium time

consumption and costs (Table 1.3). The MiSeq and HiSeq platforms were suitable in depth sequencing for metagenomics study and diversity analysis as a tool to environmental screen because of low cost and large output of target amplicons sequences (Thomas, Gilbert et al. 2012).

Table 1.3 The comparison among three predominant NGS platforms for de novo sequencing.

Feature	HiSeq2500-Highoutput	Miseq	454 Roche Pyrosequencing
Number of Reads	150-180M/lane	12-15M (v2) 20-25M (v3)	5 M
Read length	2 x 100 bp	2 x 300 bp	500 bp
Yield per lane	Up to 200 Gb	Up to 15 Gb	Up to 2 Gb
Instrument Time	~12-14 days	~2 days	7 hrs
Pricing per Gb	$59	$108	$400
Error rate	0.03	0.03	0.1

1.5. Bioinformatics analysis of high throughput data

NGS platforms output millions of bp of sequencing data and thus raise a significant challenge in metagenomic sequence analysis, requiring large amounts of data storage and management. In addition considerable issues relate to how to address and eliminate the high sequence noise rate which was generated from NGS platforms and provide efficient and rapid implementation of bioinformatics analysis on high throughput data. A number of data analysis platforms exist including Quantitative Insights into Microbial Ecology (QIIME), Mothur, WATERS, RDPipeline, VAMPS, Genboree, SnoWMan, Metagenomics RAST (MG-RAST), Uparse, and Oligotyping to

deal with pyrosequencing or MiSeq sequence data. These platforms all provide open-source bioinformatics pipelines for raw DNA sequence data analysis including several critical steps such as demultiplexing, quality filtering, sequence alignment, species clustering and classification and finally diversity analysis, statistical analysis and graphic presentation. This study aimed to compare the features and functions of QIIME and Oligotyping for the analysis of efficiency and accuracy especially with regards to diversity analysis of environmental samples.

1.5.1. QIIME

QIIME was the most well-known bioinformatics analysis platform for sequence data analysis from diverse sequencing technologies such as 454 pyrosequencing and MiSeq sequencing (Caporaso, Kuczynski et al. 2010, Gonzalez and Knight 2012). QIIME was developed with multiple functions for sequencing analysis via combination of recently published python scripts and algorithms incorporated directly into the platform (Kuczynski, Lauber et al. 2012). The features of published algorithms in QIIME were aimed to efficiently handle large amount of sequence data and reduce the time and increase efficiency.

In order to prevent sequence noise generated from NGS approaches from reducing the accuracy of species classification, quality filtering was a priority to eliminate random insertion/deletion, sequence noise, and chimeric sequences from 454 or Illumina sequencing data. Operational taxonomic units (OTU) were generated based on sequence similarity in order to reduce sequence datasets to smaller sets by clustering similar sequences together. Sequence alignment and species classification used curated reference databases including the Basic Local Alignment Search Tool

(BLAST), Ribosomal Database Project (RDP) and Greengenes databases. All the sequence data were summarised in ecological indices or statistical analyses such as heatmaps, networks, phylogenetic trees, and alpha and beta diversity metrics.

There are limits on the sequence lengths QIIME will use only a minimum of 200 bp to maximum of 1000 bp are accepted into this pipeline but this platform provides several functional and different algorithms that users can select to modify, validate or optimise this program (Nilakanta, Drews et al. 2014). Advantages of QIIME include continuous updates to incorporate the latest methods of metagenomics analysis and provision of a wide range of algorithms to perform sequence clustering, alignment, taxonomic assignment, phylogenetic tree building and diversity analysis.

1.5.2. Oligotyping and Minimum Entropy Decomposition (MED)

High throughput sequencing of a highly conserved 16S rRNA gene region was an appropriate molecular probe for investigation of whole bacterial community present in our environmental reservoirs. This probe can also monitor microbial dynamic diversity over diverse climatic change or spatial distance difference (Huber, Mark Welch et al. 2007). In order to delineate total bacterial diversity, each sequence needs to be considered. However there are two limitations in QIIME; first was that OTU are based on abundance similarity clustering. Non-dominant sequences with subtle differences will not be represented in their own OTU, but will be clustered with more dominant similar sequences. The second limitation was that a large proportion of the 16S rRNA reference sequences were isolated from culture so it offers poor resolution for environmental diversity analysis when species classification still relies on curated databases. (Garrity 2004).

The computational method of oligotyping aimed to reveal the concealed diversity in sequence analysis without comparison to current databases. Species clustering in oligotyping relies on comparing all positions in an amplicon, and using entropy analysis to select positions with appropriate levels of entropy (Figure 1.4. Step 2). The formula of entropy used in Oligotyping is listed in Equation 2.4 to achieve the effective accumulation of positions of interest to emphasise the high variation in the sequence reads. Random insertion and deletion sequence errors cause low levels of entropy as they are randomly distributed through the sequence, and are thus excluded from oligotyping analysis (Eren, Maignien et al. 2013). A previous study reported that oligotyping was able to differentiate between two high similarity bacterial species with only 0.2 % variation in the short hypervariable 16S rRNA region (Eren, Zozaya et al. 2011). Therefore Oligotyping was chosen as suitable for differentiation of SGM members especially MAC because these bacterial species have only 0.3 % difference in 16S rRNA gene region (Chaves, Sandoval et al. 2010).

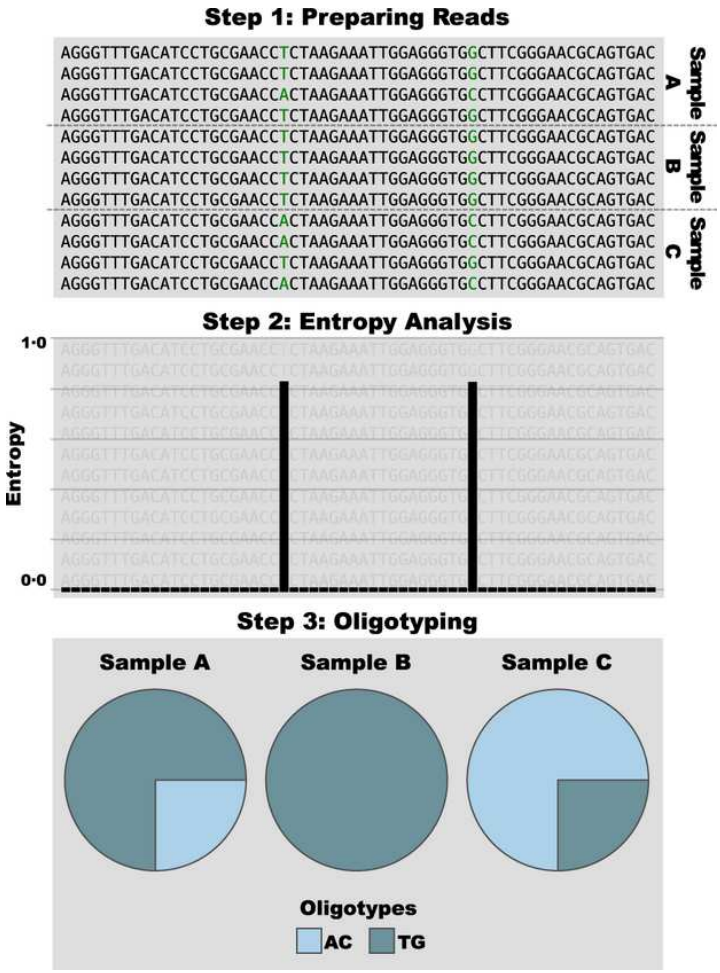


Figure 1.4 The major processes in Oligotyping analysis. The step 1 indicated that the diverse sequence reads in an OTU in the dataset. Step 2 depicted that the Shannon entropy analysis was applied on these reads to accumulate the higher variation position was revealed. Step 3, each oligotypes were identified according to entropy analysis (Eren, Maignien et al. 2013).

The species classification in Oligotyping relied on BLAST, using a curated reference database, to identify closely related sequence for taxonomic analysis. This limitation of Oligotyping resulted in being unable to address rare proportion of undefined bacterial species. However the potential of entropy analysis optimised differentiation of homogenous units through only looking at the fraction of the available nucleotide data instead of species classification based on sequence similarity, to perform directly diversity analysis (Figure 1.5) (Eren, Morrison et al. 2015). Each node generated from

MED was segregated as a new phylogenetically homogenous unit in the end of analysis. Sequences are assigned to nodes using the most entropic position, entropy within nodes is then recalculated, and new nodes are then created based on the most entopic position. This process continues until all sequences are in nodes with no entropy above a user defined level. These nodes can be used for analysis in the same way as OTU can in QIIMEs pipeline. A current study revealed that MED as a sensitive technology for diversity analysis of closely related bacterial species and this approach facilities as a biomarker in-depth metagenomics analysis (Eren, Morrison et al. 2015).

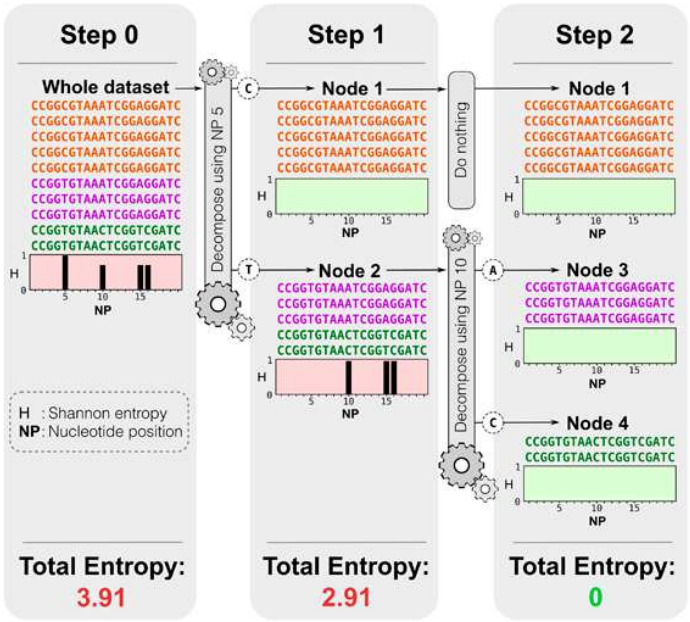


Figure 1.5 Flow chart illustrates the principle of MED to decompose each high entropy sequence location to different Node group. The step 0 Shannon entropy analysis was performed in Oligotyping. In step 1, any node with a Shannon entropy more than 0 was decomposed and built a reanalysis for regeneration of new nodes in step 2 (Eren, Morrison et al. 2015).

A comparison was performed of two bioinformatics NGS analysis pipelines and their associated algorithms consisting of Uparse and MED (Table 1.4). Only QIIME was limited by the sequence read length but the QIIME platform provided several

algorithms supporting oligotyping analysis methods such as denoising, chimera removal, alignment and statistical analysis. Uparse is an independent programme developed to improve QIIME with high accuracy and sensitivity of OTU generation and with its own denoising, chimera removal, and alignment system but the diversity and statistical analysis still relied on QIIME (Edgar 2013). Oligotyping and MED analysis pathways were found to be superior to QIIME and Uparse in terms of diversity and statistical analysis.

Table 1.4 The comparison of two sequencing analysis methods, QIIME and Oligotyping, with their associated algorithms.

	QIIME	Uparse	Oligotyping	MED
Sequencing platform	454 Pyroseq & Illumina	454 Pyroseq & Illumina	454 Pyroseq & Illumina	454 Pyroseq & Illumina
Read length	200 ~ 1000 bp	0 ~ user defined	0 ~ user defined	0 ~ user defined
Denoising	Denoiser	Usearch	Denoiser, Usearch	Denoiser, Usearch
Chimera removal	ChimeraSlayer, Usearch	Usearch	ChimeraSlayer, Usearch	ChimeraSlayer, Usearch
Sequence cluster	OTU	OTU based usearch	Shannon entropy algorithm	Shannon entropy based Node algorithm
Species classification	BLAST, RDP, Greengenes	BLAST, RDP, Greengenes	BLAST	None
Alignment	Pynast, Mothur	Usearch	Pynast, Mothur	Pynast, Mothur
Phylogenetic Tree	Phylogenetic Tree	Phylogenetic Tree	Clustering Tree	Clustering Tree
Diversity analysis	Alpha & Beta diversity	Alpha & Beta diversity	Clustering & NMDs	Clustering & NMDs
Statistical mapping analysis	PCoA, NMDS and ANOSIM etc.	PCoA, NMDS and ANOSIM etc.	NMDs and ANOSIM etc.	NMDs and ANOSIM etc.

1.6. Spatial study and statistical analysis

It is not possible to delineate the composition of the whole bacterial community in the ecosystem via identification of bacterial species using qPCR with specific primers.

Whole environmental screening was necessary with diversity analysis; however cultivation is not an acceptable tool, with only a small proportion (1%) of bacteria being culturable from soil (82). In addition inaccurate and insensitive estimation of species richness have been obtained from cultivation techniques due to insufficient coverage of the bacterial population, and ambiguous selective media culture. The remedy to this predicament was to provide accurate and robust methods for metagenomics analysis from environmental samples.

The spatial study consisted of a wide range of geographical bioinformation systems and spatial statistical analyses which monitor the correlation of subjects within environments, bioclimatic change or geographical distance. The R program has a number of packages suitable for analysis of spatial studies, for example ggplot2 and ggmap. The application of R was also capable of statistical techniques including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, and clustering.

Several statistical analyses were used in this study to analyse the relationship of target bacterial species and communities associated with diverse sample sets and seasonal changes. Analysis of variance (ANOVA) was used when data was normally distributed, while the Mann-Whitney test was used on non-normally distributed datasets. In contrast to single variable analysis, multivariate analyses such as hierarchical clustering for multiple variables in an observation were based on measures of dissimilarity within each cluster (Boratyn, Datta et al. 2006). In addition, mapping statistical analyses also used measures of dissimilarity within sample sets; Principle Coordinate Analysis (PCoA) plots were made using QIIME to delineate the

relationship of beta diversity between sample clusters in three dimensions, and Non-metric Multidimensional Scaling (NMDS) plots derived from R programme and were made for visualisation of correlation within sample sets based on ranking distance dissimilarity analysis. The centroid of each ellipse referred to the mean of each group and the shape of each group was defined by the group covariance. Analysis of similarities (ANOSIM) is a non-parametric statistical analysis based on Bray-Curtis measures of similarity to identify difference between two or more multivariate groups based on any disease measure (Clarke 1993).

1.7. Case study in Tanzania

Tanzania has a population of approximately 50 million. The country was reported as a high-TB-burden country as new TB incidence was 160 per 100,000 population annually, but additionally 48 % of TB patients have extra-pulmonary disease (Zumla, George et al. 2015). Other issues were HIV associated TB and multidrug-resistant TB (MDR-TB) within TB patients in Tanzania. The proportion of TB patients with known HIV status was 83% as recorded and 10% of TB cases were notified as MDR-TB cases (Zumla, George et al. 2015).

Extra-pulmonary TB disease is considered a serious and public health burden in Tanzania but this aspect of the disease resulted from several factors consisting of HIV associated atypical TB infection and cross-species TB infection. The causal agent of well-known cross-species TB infection was Mb, this animal pathogen makes up a large proportion of extra-pulmonary TB cases at 53 % compared to only 18 % caused by

Mtb (Katale, Mbugi et al. 2012). The cross-species TB infection caused by Mb reported in Tanzania was 17 % of TB cases (Tanzania 2006). It was mentioned that the treatment and diagnosis of typical and atypical TB cases was a priority but it relied upon the tuberculin skin test and specimen smear culture technique which were less sensitive and more time consuming for urgent cases especially in Tanzania. In addition as poverty and malnutrition are risk factors and occur in Tanzania, this was directly associated with the development of TB prevalence and incidence.

The National Institute of Health (NIH) funded this project for a case control study of TB infection between humans and animals according to development of a rapid and sensitive technique for treatment and diagnosis of typical and atypical TB cases. The University of Warwick is a part of academic collaboration between organisations: University of California, Davis (UCD) and San Francisco (UCSF) in the USA, Sokoine University of Agriculture (SUA) and the National Institute for Medical Research in Tanzania. Warwick is responsible for the environmental detection and analysis which is integrated with the human (UCSF) and animal (UCD) study.

1.8. Aims and Hypotheses

General hypotheses for this project

Hypothesis 1: In areas of high endemic disease cattle will be a major source of Mb; with exposure risk to pastoralists. Faecal shedding provides a useful monitor of animal disease status regardless of season.

Hypothesis 2: Environmental shedding by TB infected humans provided a useful indicator of disease status and correlates with known clinical status.

Hypothesis 3: Cattle are primary host for Mb and this is main cause of bTB so herds rarely shed other mycobacterial pathogens.

The specific aims and objectives for environmental component of the project are as follows:

Aim 1: Determine the environmental reservoir of Mb and Mtb in household dust, soil and livestock faeces (cattle and goat) in the target area within the Ruaha region. (Figure 1.6 wet work)

Objective 1: Develop and test molecular methods for detection and quantification of Mtb and Mb in environmental samples.

Objective 2: Use optimal methods for Mb and Mtb to quantify by qPCR and IMC environmental reservoirs in Ruaha dust, faeces, soil and water.

Aim 2: Compare Mb and Mtb load in samples taken in dry and wet seasons in areas of different infection levels in Tanzania. (Figure 1.6 dry work with statistical analysis and spatial study)

Objective 3: Collate data on infection in human and animal populations and relate this to environmental reservoirs to determine correlations and establish environmental hotspots for disease transmission.

Aim 3: Determine diversity of SGM communities in faeces and soil to provide quantitative data for future disease model development. (Figure 1.6 dry work with informatics analysis)

Objective 4: Focus on correlation of each sample set in environmental hotspots and monitor of mycobacterial isolates in clinical samples.

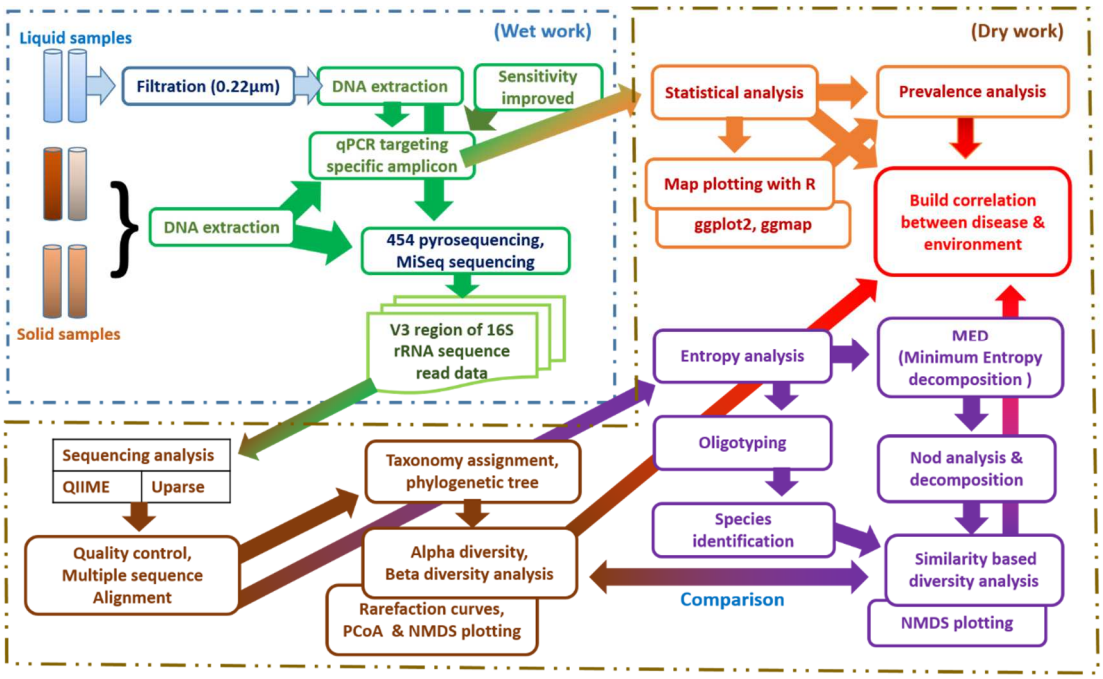


Figure 1.6 Flow chart of my PhD project including wet and dry works.

Chapter 2 Materials & Methods

2.1. Establishment of field sites

2.1.1 NIH study A

The aim of the NIH study was to understand the correlation between TB patients and environment surrounding their living space including dust from food preparation area, bathroom facilities and livestock housing area in domiciles within Ruaha region in the Iringa administrative region (7°46'S, 35°42'E) of Tanzania (Figure 2.1 and Figure 2.2). Historical data was available and tuberculosis disease data was collected from National Institutes of Health (NIH) project and World Health Organisation (WHO). The NIH sponsored data was from the HALI project which indicated the prevalence and incidence of tuberculosis, HIV-TB co-infected case, extra-pulmonary and atypical tuberculosis incidence caused by non-MTBC member pathogen in Ruaha area for the years 2006 and 2007 (P 2011).

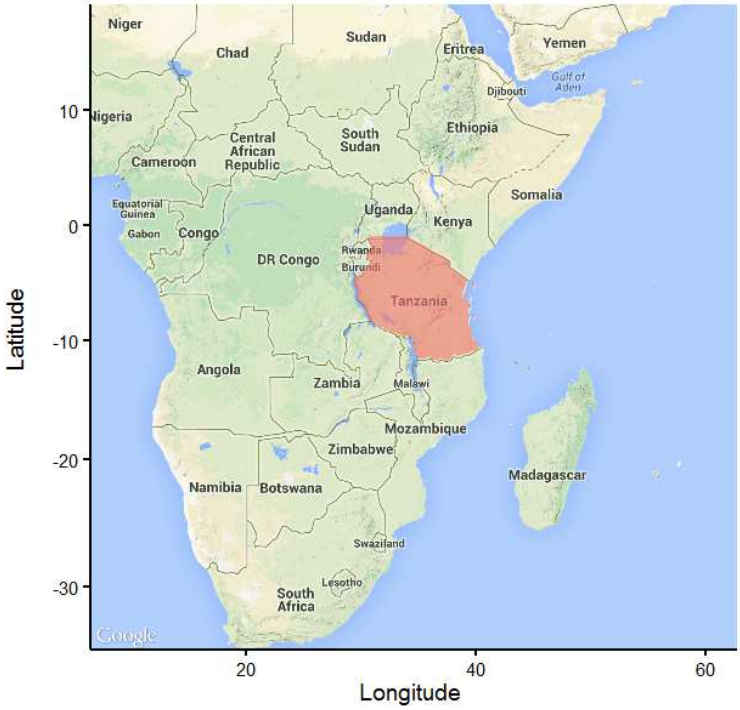


Figure 2.1 Map of Africa continent to show the location of Tanzania highlighted in red

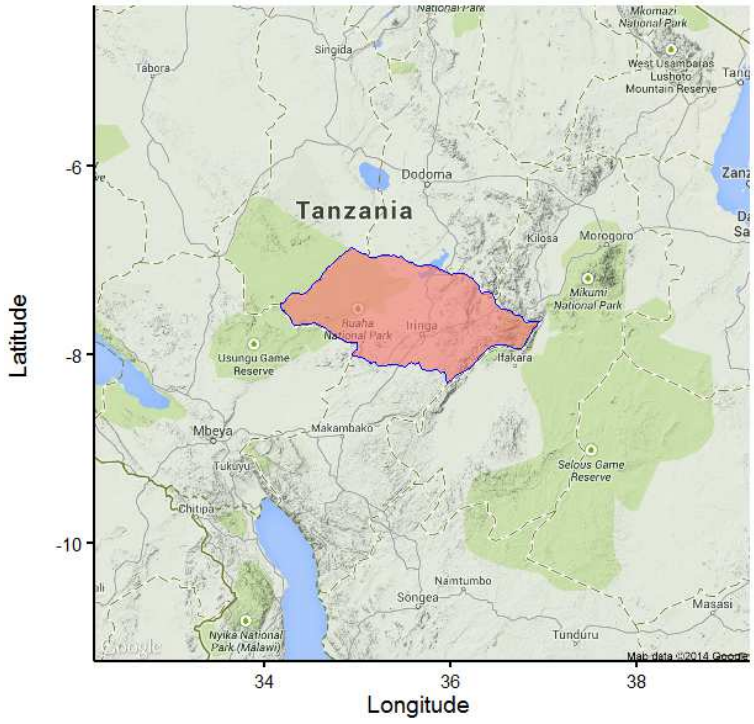


Figure 2.2 Map of Tanzania to show administrative area of Iringa highlighted in red. This area included east region of the Ruaha national park

Two sampling approaches were used;

1. To collect information from TB patient incidents used in the National Institutes of Health project. This was passive case finding through the self-reporting of TB patients to hospitals according to their symptoms and then sputum was collected and cultures obtained by the investigated teams in the National Tuberculosis and Leprosy Programme in Dar es Salaam, Tanzania.

2. The second one was active case finding via hiring and training local people in specific areas in two high potential and TB prevalent region, Pawaga and Idodi districts in the Ruaha region and help to refer household contacts suspected of having tuberculosis to seek following screening and identifying to the local clinic.

Patients identified as TB positive cases who agreed to participate in the study, were quizzed for clinical and epidemiologic data entered in the NTPL reporting form. The coordinates of their house was recorded and mapped within the interview and agreed between TB patients and NTPL training coordinator. The details of the process (Table 2.1) showed the households with at least one TB patients was identified as TB case household. In addition, the case-control study was introduced in this part of the investigation, control household was identified as no TB patients in the house by the NIH project but location matched based on district.

There were 185 households in total with 100 TB case households and 85 matched control households. Dust was collected form these households during two years during 2012 and 2013 with the aim of investigation environmental shedding of Mtb and Mb in the Iringa administrative region (7°46’S, 35°42’E) of Tanzania.

Table 2.1 The process of TB case identification with a thousand TB case

Culture positive TB case identified (n=1000)

↓

Index case interview

- Index case interviewed with standard interview form by study coordinator with one week of diagnosis
- Clinical, demographic and epidemiologic data collected or extrapolated from existing chart
- Roster of household contacts (names and ages) collected from index case

↓

Contact Investigation

- Contact investigator goes to house of index case, screens contacts with standard interview form
- Any contact suspect of TB asked to go to clinic free of charge for TB sputum exam
- HIV testing offered to all cases and contacts

↓

Collection and processing of clinical isolates

- Clinical isolates will be collected for all index and secondary cases per standard clinical procedures
- Molecular genotyping, drug susceptibility testing will be performed at NIMR and UCSF

2.1.2 Study of disease reservoirs in the environment of pastoralists and their livestock study B.

Different types of samples were collected from the environmental reservoir including livestock faeces, soil samples from livestock enclosure (boma as described in many parts of Africa which is a mixture of bare earth and animal manure in animal enclosures), household dust, water and sediment from six villages distributed in Ruaha region in the Iringa administrative region (7°46'S, 35°42'E) of Tanzania (Figure 2.2.). Samples were collected once during each season depending on precipitation and average rainfall days within the month (Figure 2.3) to be identified as wet and dry season. There were two sampling occasions during this study to compare the potential risk between extremely different climatic conditions of the wet and dry seasons. The samples were collected by the Warwick team in September of 2012 describe as dry season and February of 2014 as wet season.

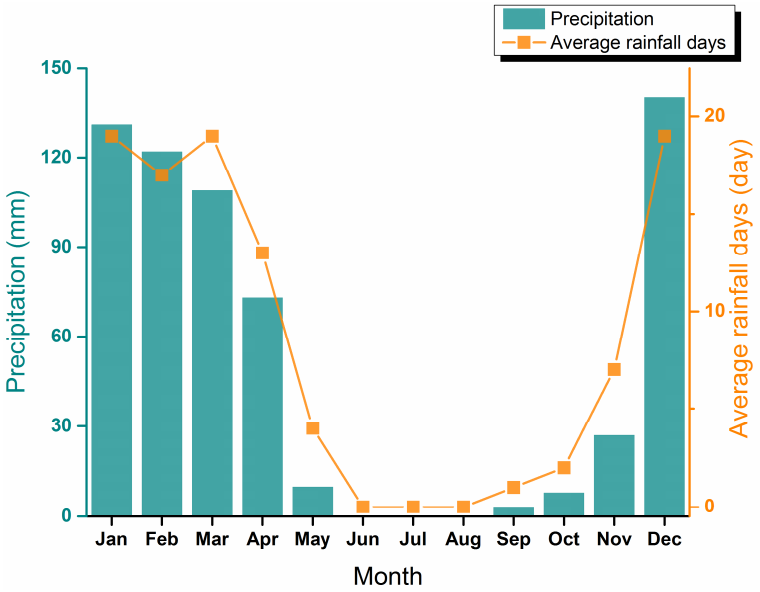


Figure 2.3 The precipitation and rainfall days within a month in the Iringa administrative region.

Environmental samples were collected from five households we selected from each village located near the Ruaha national park (7°30'S, 35°0'E). The park was established in 1964 and is the largest national park in Tanzania with a large population of wildlife animals. The Ruaha River is the main water source supporting wildlife, livestock and pastoralists living nearby. Proximity to the park was therefore increasing the risk for cross contamination between wildlife and livestock while they shared the same water source. Six pastoralist villages close to the park were selected to investigate the correlation between animal disease and environment. Six villages were located between the Longitude 35°04'E to 35°35'E and Latitude 7°16'S to 7°55'S.

Villages were identified from 1-6 in chronological order of investigation in the wet and dry seasons and coordinates recorded using the Garmin eTrex® 10 handheld GPS device (Garmin (Europe) Ltd, Southampton, UK). We selected five households from each village to compare disease shedding from livestock and environmental

dispersion of the Mb and Mtb pathogens using qPCR. In this study, households belonging to the same villages were represented by the household number followed by the village number.

Water and sediment samples were chosen from the main or branch river Ruaha close to the pastoralist villages and seven places selected in dry season compared to eight places in wet season because of water scarcity (Tungamalenga-Ilenga) in the dry season we were unable to collect samples. Water sampling location was between the Longitude 35°04’E to 35°33’E and Latitude 7°16’S to 7°56’S.

Table 2.2 The detail information of season sample collection

Village	Name of Household	Village labelled	Region	Season samples collection	House ID	Altitude (m)	Latitude	Longitude	Family member (No. of Adult\children)	Supplemental information H I	Supplemental information H II	Own cattle (No.)	Supplemental information A I	Own goats (No.)	Supplemental information A II
Malinzanga	Mzee posraui	1st	Middle	Wet and Dry	H1-1	867	-7.608444	35.354194	11\17	Y (2 in 2012 and 2 in 2013)	TB Case household	Y	110	Y	140
Malinzanga	Serendu	1st	Middle	Wet and Dry	H1-2	862	-7.604417	35.356306	10\17	No	Control household	Y	70	Y	66
Malinzanga	Petrokwayesa	1st	Middle	Wet and Dry	H1-3	905	-7.634278	35.377833	7\4	NI	N	Y	40	Y	150
Malinzanga	Lopina Sanago	1st	Middle	Wet and Dry	H1-4	875	-7.630389	35.377472	4\16	NI	N	Y	25	Y	40
Malinzanga	Julius Lulandala	1st	Middle	Wet and Dry	H1-5	892	-7.639444	35.360306	5\12	NI	N	Y	80	Y	50
Itunundu	Mzee Shija	2nd	PAWAGA	Wet and Dry	H2-1	740	-7.357722	35.526778	41\20	NI	N	Y	150	Y	300
Itunundu	Mzee Hetee Lunga	2nd	PAWAGA	Wet and Dry	H2-2	743	-7.360889	35.528611	16\15	NI	N	Y	300	Y	30
Itunundu	Mzee Chagululo	2nd	PAWAGA	Wet and Dry	H2-3	742	-7.359583	35.528167	6\2	NI	N	Y	50	Y	20
Itunundu	Lyochi Mejia	2nd	PAWAGA	Wet and Dry	H2-4	745	-7.363917	35.535444	9\16	NI	N	Y	40	Y	50
Itunundu	Kuiwa Makelemo	2nd	PAWAGA	Wet and Dry	H2-5	754	-7.362444	35.535306	2\4	NI	N	Y	50	Y	17
Mboliboli	Mzee Ndaki	3rd	PAWAGA	Dry only	H3-1	733	-7.280694	35.582806	7\4	NI	N	Y	30	Y	20
Mboliboli	Paulo Garobako	3rd	PAWAGA	Dry only	H3-2	730	-7.278583	35.581194	12\2	NI	N	Y	40	Y	28
Mboliboli	Kilo Moja	3rd	PAWAGA	Dry only	H3-3	716	-7.283806	35.563556	14\5	NI	N	Y	60	Y	15
Mboliboli	Paulo digaa	3rd	PAWAGA	Dry only	H3-4	714	-7.283111	35.569	8\2	NI	N	Y	40	Y	30
Mboliboli	Ihelaga Paulo	3rd	PAWAGA	Dry only	H3-5	715	-7.288222	35.567083	10\4	NI	N	Y	45	Y	40
Kinyika	Musa Mabula	4th	PAWAGA	Wet and Dry	H4-1	739	-7.360167	35.448611	10\10	NI	N	Y	100	Y	125
Kinyika	Elias Mabala	4th	PAWAGA	Wet and Dry	H4-2	741	-7.358556	35.448028	8\4	NI	N	Y	100	Y	20
Kinyika	Sungiya Maasaw	4th	PAWAGA	Wet and Dry	H4-3	739	-7.361722	35.45	2\2	NI	N	Y	60	Y	50
Kinyika	Juma Somei	4th	PAWAGA	Wet and Dry	H4-4	739	-7.363389	35.450556	4\9	NI	N	Y	20	Y	30
Kinyika	Kimalu Tekelu	4th	PAWAGA	Wet and Dry	H4-5	747	-7.365333	35.44975	4\6	NI	N	Y	5	N	0
Kitis	Elias Kichianga	5th	IDODI	Wet and Dry	H5-1	937	-7.807222	35.0845	3\7	NI	N	Y	15	Y	10
Kitis	Uichianga Nai	5th	IDODI	Wet and Dry	H5-2	942	-7.808306	35.083972	5\5	NI	N	Y	20	Y	30
Kitis	Mzee Philipo Miwani	5th	IDODI	Wet and Dry	H5-3	942	-7.805028	35.092556	20\10	NI	N	Y	100	Y	200
Kitis	Didoyi Philipo	5th	IDODI	Wet and Dry	H5-4	942	-7.807889	35.091861	2\2	NI	N	Y	20	N	0
Kitis	Paschali Gweldemo	5th	IDODI	Wet and Dry	H5-5	929	-7.793861	35.1175	2\5	NI	N	Y	35	Y	50
Tungamalenga	Mangunda Kuyca	6th	IDODI	Wet and Dry	H6-1	983	-7.865861	35.141278	9\8	NI	N	Y	47	Y	20
Tungamalenga	Athmani Lolayo	6th	IDODI	Wet and Dry	H6-2	971	-7.863556	35.143389	3\7	NI	N	Y	20	Y	20
Tungamalenga	Dotto Shangai	6th	IDODI	Wet and Dry	H6-3	964	-7.859833	35.145472	3\9	NI	N	Y	20	Y	30
Tungamalenga	Zuberi Kateyi	6th	IDODI	Wet and Dry	H6-4	959	-7.861722	35.1475	6\2	NI	N	Y	17	Y	18
Tungamalenga	Manywele	6th	IDODI	Wet and Dry	H6-5	963	-7.862972	35.147222	6\4	NI	N	Y	97	Y	59

Supplemental information H I: Number of TB patients in this household

Supplemental information H II: This household has been reported in NIH project

Supplemental information A I: The number of cattle

Supplemental information A II: The numbers of goats.

Table 2.3 The detail information of season water sample collection

Name of River	River ID	season sample collection	Altitude	Latitude	Longitude	river condition	flow rate (m/s)	river width	river depth	pH water hole (river)
Ikonogo	IKO	Both wet and dry	891	-7.63864	35.37211	Stagnant	0	30	3	9
Savani	SAV	Both wet and dry	739	-7.272	35.53858	slow flow	2	30	2	9
Ikwavila	IKW	Both wet and dry	1123	-7.41117	35.47775	Fast flow	0.5	3	0.5	9
Makife	MAK	Both wet and dry	1014	-7.92797	35.09736	slow flow	1	15	2	9
Tungamalenga	TUN	Both wet and dry	980	-7.89358	35.07006	Fast flow	0.5	30	1	9
Tungamalenga - Ilenga	TUI	Wet only	964	-7.83544	35.08503	slow flow	1	20	2	9
Itunundu	ITU	Both wet and dry	743	-7.34531	35.50122	Fast flow	0.5	20	0.5	9
Idodi	IDO	Wet only	938	-7.82169	35.09342	Fast flow	0.6	10	0.5	9

2.1.3 Climate data collection

Temperature ranged between 28 to 32˚C with an average high temperature of a year and 15 to 20˚C with an average low temperature of a year (Figure 2.4). The precipitation and average rainfall days; however, can be separated as wet and dry seasons. The dry season is between May to November and the wet season is December to April. There were > 10 days rainfall in wet season and < 10 days in dry season, and < one wet day in June, July, August and September. Over 110mm precipitation fell within the month in the wet season compared to only 6mm on average in the dry season (Figure 2.3).

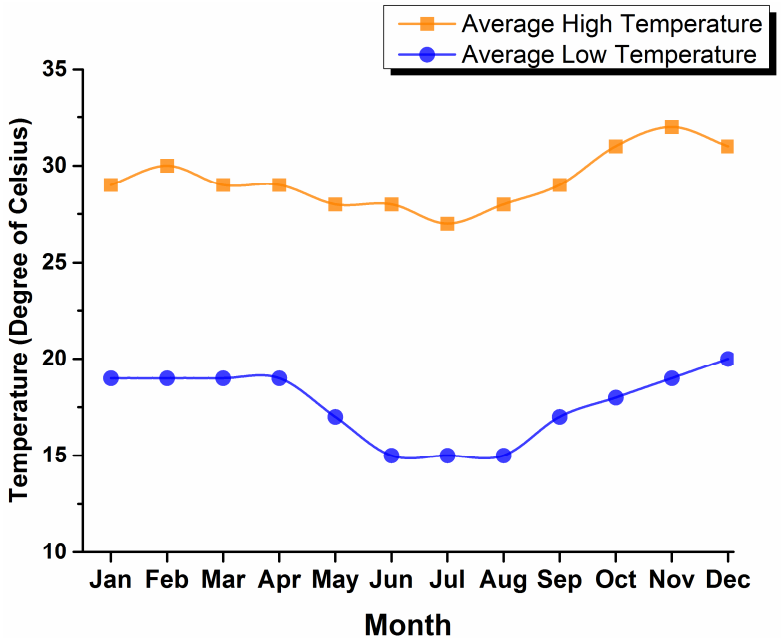


Figure 2.4 The average monthly temperature in Iringa administrative region.

2.2 Environmental sample collection

2.2.1 Environmental sample collection from Study A, TB case and control households and Study B, rural village houses and livestock

Study A: the household dust for the NIH project was household dust only was collected by the NIMR/SUA integrated team.

Study B: Collection of the environmental samples chosen in this study were faeces from livestock including cattle and goat, boma soil, dust from households, water and sediment sample from selected water sources.

For household dust collection, the top 3 cm of dust was removed from food preparation area, washing facilities and livestock housing area in three different domiciles including their first house with food preparation area and washing facilities, second house is a chamber for men as well as third one for women and children

(Figure 2.5). The six dust samples were collected in each house including three from inside of house and another three from outside, using the sterilised spatula into the plastic sample bags for each household and pooled immediately to make a mixed sample to represent each individual household. The bag was shaken and three biological replicates of 2 g selected in the Eppendorf tubes and stored at 4 ˚C. All samples were labelled and blinded for shipment to the UK.

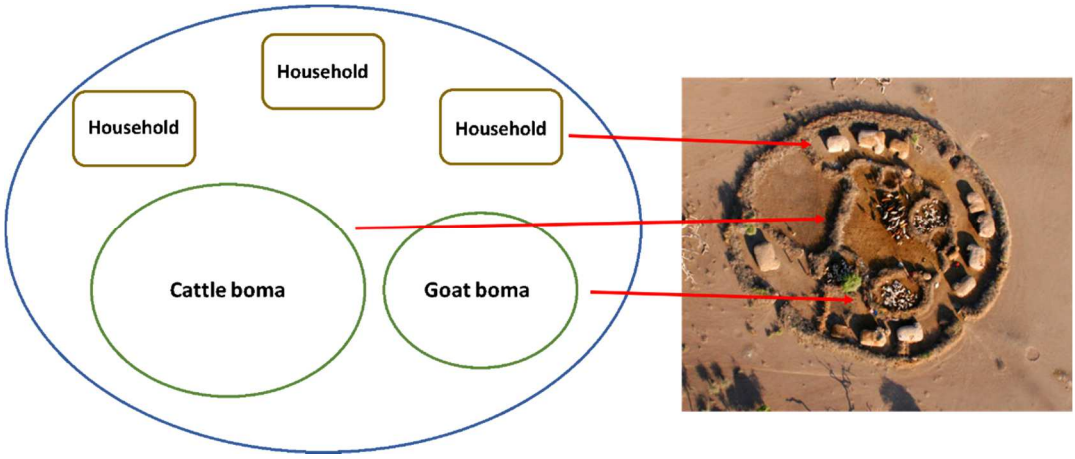


Figure 2.5 Sampling schematic of boma soil and household dust collection points per household. The picture on the right was satellite image to show the structure of household. (http://enduimet.org/wp-content/uploads/2013/05/Maasai-boma.jpg)

2.2.2 Sample collection from pastoral homesteads of dung and boma soil

For environmental sampling the number of replicate samples required from a given location to ensure a < 1% probability of a false negative will be calculated in line with our previous protocols in the UK (Courtenay, Reilly et al. 2006), calculated by equation 2.1.

$$\Psi_N = (1-p)^m$$ **equation 2.1.**

Where Ψ_N is the probability that an Mb positive site tests negative (i.e. a false negative), p is the probability of a positive site testing positive, and m is the number

of samples tested per site. Required values of m from our published studies in the UK were between 6 and 8 to give $\Psi_N < 0.01$. We evaluated spatial and temporal factors to calculate *m* for our study populations, and conducted a pilot study to determine the value of *p* by testing 15 replicate samples from 10 spatially independent sites in low and high incidence regions.

The top 1 cm^3 of dried faeces exposed to the sun was removed prior to sample collection to reduce the effect of UV irradiation. At each of site depending on number of livestock and size of livestock enclosure, ten faecal samples were collected from cattle faeces and four from goat faeces, approximately 2 g faeces were collected using sterilised spatula into the Eppendorf and stored in cooled boxes with an ice pack at 4 ˚C. The same procedure was applied on the boma soil collection but the nine soil samples were selected in each cattle and goat boma area. The soil samples were collected using the sterilised spatula into the plastic sample bags from each boma area and pooled immediately to make a mixed sample to represent individual homestead. The bag was shaken and three biological replicates of 2 g selected in the Eppendorf tubes and stored at 4 ˚C. These sample were deep frozen using liquid nitrogen then shipped to the UK in a dry shipper under appropriate biohazards labelling UN3373.

The moisture of soil was detected using the Tecpel pH 707 Soil and Moisture Tester (Tecpel, Taipei, Taiwan) for 10 min equilibration after inserting in the soil around 5 cm deep from surface. The range of moisture content in the cattle and goat boma soil in the dry season was compared to the wet season (Figure 4.13)

2.2.3 Other environmental samples: water and sediment samples

The water sample was collected from the surface of running water facing upstream into the current and > 500 ml volume was collected from each sampling site close to villages and each 250 ml water was filtered using 100 ml sterilised plastic syringe and MicrofilV filtration device (Thermo Fisher Scientific, Leicestershire, UK) with 0.22 µm mixed cellulose esters white gridded filters (Millipore, MA, USA). After collection the filters were removed from the plastic holder using sterile forceps and air-dried. The filters were then rolled and folded and stored in 2 ml Eppendorf in a cool box with ice packs.

Representative mixtures of sediment consisting of sedimentary rock, mud and sand were collected using the sterilised plastic spatula scooped along the bottom of surface river body < 1.5 M deep in the upstream direction. The sediment sample was placed into the 50 ml sterilised plastic universal tube (Scientific Laboratory Supplies Ltd, Nottingham, UK) and homogenised. Excess water was removed from the container prior to storage.

The pH value for the river was tested using New Hydrion® pH indicator strips (Micro Essential Laboratory, New York, USA) (Table 2.3).

All samples were deep frozen using liquid nitrogen then shipped to the UK in a dry shipper under appropriate biohazards labelling UN3373.

2.2.4. 16S rRNA sequencing results from livestock and wildlife animal in Tanzania

The livestock lesion samples from slaughterhouse compared to wild buffalo tissue samples from Ruaha national park were tested using 16S rRNA sequencing at Sokoine

University of Agriculture in Tanzania (Table 2.4). *M. intracellulare* was detected in tissue samples of wildlife compare to *M. lentiflavum* which present in most of the wild buffalo and cattle lesions samples. These two SGM species belongs to opportunistic mycobacteria (Mwikuma, Kwenda et al. 2015) and can be identified in drinking water supplies as a possible source of NTM infection in humans (Falkinham, Norton et al. 2001, Marshall, Carter et al. 2011).

Table 2.4 Comparison of different species identification results from livestock lesion (A) and wild buffalo tissue (W) using 16S rRNA sequencing at Sokoine University of Agriculture in Tanzania.

Strain number (Tanz)	Strain number (FLI)	Species
W11407	12MA1273	***M. indicus pranii / M. intracellulare***
W11398	12MA1275	***M. indicus pranii/ M. intracellulare***
W1104	12MA1276	***M. indicus pranii/M. intracellulare /M. arosiense***
W11356	12MA1277	***M. vulneris/ M. intracellulare***
W11413	12MA1279	***M. lentiflavum***
W11363	12MA1281	***M. lentiflavum***
A004LN	12MA1285	***M. lentiflavum***
A006LN	12MA1286	***M. intracellulare / M. indicus pranii***
W11411	12MA1274	***M. lentflavum/palustre/simiae***
A001LN	12MA1282	***M. lentiflavum***
A004LN	12MA1284	***M. lentiflavum/palustre/simiae***
A010	12MA1289	***M. lentiflavum/palustre/simiae***
A007LN	12MA1290	***M. lentiflavum***

2.3. Bacterial species and growth condition

The bacterial species used in this study are listed in Table 2.5. All manipulations were conducted in the category III level laboratory. The stocks were resuscitated on the Difco™ Middlebrook 7H10 agar (BD, Oxford, UK) and sealed with the Parafilm® M

Barrier Film (SPI Supplies, West Chester, USA) and proliferated at 37 °C . A single colony was selected and inoculated onto the Difco™ Middlebrook 7H9 broth with 10 % Middlebrook ADC enrichment (BD, Oxford, UK) contained and sealed with the Parafilm® M Barrier Film at 37 °C in the incubator shaking at 150 rpm.

Table 2.5 Mycobacterium species strains used in this study

Taxonomy	strain	Growth rate
Mycobacterium bovis BCG ATCC	Pasteur	Slow
Mycobacterium bovis ATCC		Slow
Mycobacterium tuberculosis ATCC		Slow

2.4. DNA extraction

Different extraction commercial kits consisted of Blood & Tissue kit (QIAGEN, Ltd., Crawley, UK), FastDNA™ SPIN Kit for Soil (MP Biomedicals, UK), PowerFood and PowerWater® DNA Isolation Kit (Cambio, Cambridge, UK) were applied for different sample sources.

The total number of faeces samples in this study is 976 (510 from dry season and 466 from wet season), the boma soil totals 313 (171 from dry season and 142 from wet season), 165 household dust (90 from dry and 75 from wet) and 34 sediment samples (16 from dry and 18 from wet) from 30 households belonging to 6 villages in dry season and 25 households distributed in 5 villages in wet season (Table 4.1). Total DNA was extracted from 0.2 g (±0.01 g) of solid samples consisted of faeces, soil, dust

and sediment using FastDNA SPIN kit for Soil following manufacture's instruction. In summary for DNA extraction, 0.2 g sample was added to the lysis matrix tube with 1.4 mm ceramic spheres 0.1 mm silica spheres and 4 mm glass beads which was suitable for soil, sediment and faecal sample as manufacture's recommendation. Lysis buffer and sodium phosphate buffer were subsequently added into the tube for physical and mechanical disruption using Ribolyser Instrument Precellys 24 (Stretton Scientific Ltd, Stretton, UK). Removal of protein and other materials with protein precipitation buffer was performed followed by purification of DNA with ethanol using a centrifugation step was supplied until the end of DNA extraction and collection.

The total number of water samples is 34 filters consisting of 16 from 7 sample sites collected during dry season and 18 from 8 sites in wet season. DNA was extracted from 0.22 µm filter using PowerWater® DNA Isolation Kit following the manufacture's instruction. In summary, both chemical and mechanical lysis were employed on cell disruption and removal of protein, cell debris and other organic or any inorganic materials except DNA. The extracts were subjected to high salt solution and then washed with ethanol for purification and consequently DNA was eluted from the membrane using elution buffer prior to storage at -20 °C.

2.5. Real-time quantitative PCR

2.5.1. qPCR assay for MTBC

Primers targeting 16S rRNA and other genes were used in this study. For example, APTK primer was selected for targeting SGM species because their long helix 18 in V3

region with few exceptions such as *Mycobacterium lentiflavum* (Khera 2012). The published primer for Mtb targeted the, ninth region of difference (RD9)(Halse, Escuyer et al. 2011), and for Mb, the forth region of difference (RD4 scar) (Sweeney, Courtenay et al. 2007) were selected for targeting Mtb and Mb respectively. Furthermore, LepA primers were used on all members of MTBC and only Wbbl1 primer on Mtb, *M. canettii* and *M. africanum* was applied (Reddington, O'Grady et al. 2011) for comparison of sensitivity of RD9 and RD4 scar primer on screening liquid environmental samples (Figure 2.6).

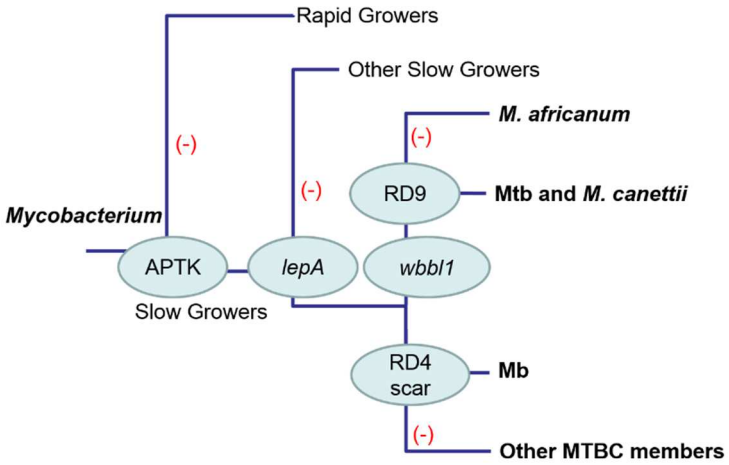


Figure 2.6 The phylogenetic tree of MTBC members and each primer highlighted in light blue oval to show the specificity of each primer targeted in this study

The qPCR detection for the primer RD9 and RD4 scar was set up using 900 nM of each primer consisting the forward and reverse primer and 250 nM probe in final concentration of working reagent, 1 mg ml^{-1} bovine serum albumen, 12.5 µl of TaqMan Environmental Master Mix 2.0 (Applied Biosystems Inc., CA, USA), 10 µl of template DNA and made up to total 25 µl with sterilised molecular grade water (Sigma-Aldrich Company Ltd, Dorset, UK). The PCR running cycles in this study was

50.0 °C for 2 min followed by 95.0 °C for 10 min then 50 cycles of 95.0 °C for 15 sec and 58.0 °C for 1 min until the end of programme.

In comparison to the RD9 and RD4 scar primer, the same reagent and qPCR programme condition was applied to the LepA and Wbbl1 primer. The exception of the final temperature of qPCR was changed from 58.0 °C to 60.0 °C according to the primer and probe difference. All primers and probes sequence used in this study (Table 2.6).

Table 2.6 Primers and probes used for the MTBC, Mtb and Mb species in this study for qPCR, internal control and diversity analysis

Primer Name	Sequence	Target group	Target gene	Reference
APTK Fw	GCT TAA CAC ATG CAA GTC GAA CGG AAA GG	SGM	405–438 bp – V1–V3	(Pontiroli, Khera et al. 2013)
APTK Rv	GTC AAT CCG AGA GAA CCC GGA CCT TCG TCG	SGM		(Pontiroli, Khera et al. 2013)
APTK Fw-Pyro	GTC AAT CCG AGA GAA CCC GGA C	SGM	405–438 bp – V1–V3	(Pontiroli, Khera et al. 2013)
APTK Rv-Pyro	GCT TAA CAC ATG CAA GTC GAA CG	SGM		(Pontiroli, Khera et al. 2013)
APTK Fw-MiSeq	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG GCT TAA CAC ATG CAA GTC GAA CGG AAA GG	SGM	405–438 bp – V1–V3	This thesis
APTK Rv-MiSeq	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGT CAA TCC GAG AGA ACC CGG ACC TTC GTC G	SGM		This thesis
RD9 Fw	TGCGGGCGGACAACTC	Mtb and *M. canettii*	RD9 region	(Halse, Escuyer et al. 2011)
RD9 Rv	CACTGCGGTCGGCATTG	Mtb and *M. canettii*	RD9 region	(Halse, Escuyer et al. 2011)
RD9 probe	6FAM-AGGTTTCA+CCTTCGAC+CC—BBQ	Mtb and *M. canettii*	RD9 region	(Halse, Escuyer et al. 2011)
RD4 scar Fw	TGTGAATTCATACAAGCCGTAGTCG	Mb	RD4 deletion region	(Sweeney, Courtenay et al. 2007)
RD4 scar Rv	CCCGTAGCGTTACTGAGAAATTGC	Mb	RD4 deletion region	(Sweeney, Courtenay et al. 2007)
RD4 scar probe	6FAM-AGCGCAACACTCTTGGAGTGGCCTAC—TMR	Mb	RD4 deletion region	(Sweeney, Courtenay et al. 2007)
LepA Fw	AGA CCG TGC GGA TCT TG	MTBC	*LepA* gene	(Reddington, O'Grady et al. 2011)
LepA Rv	CAT GGA GAT CAC CCG TGA	MTBC	*LepA* gene	(Reddington, O'Grady et al. 2011)
LepA probe	HEX-ACGGATTGGTCACCCGGATT-BHQ1	MTBC	*LepA* gene	(Reddington, O'Grady et al. 2011)

Wbbl1 Fw	**TACCAGCTTCAGCTTTCCGT**	**Mtb, *M. canettii* and *M. africanum***	***Wbbl1* gene**	**(Reddington, O'Grady et al. 2011)**
Wbbl1 Rv	**GCACCTATATCTTCTTAGCCG**	**Mtb, *M. canettii and M. africanum***	***Wbbl1* gene**	**(Reddington, O'Grady et al. 2011)**
Wbbl1 probe	**FAM-ATGGTGCGCAGTTCACTGC-BHQ1**	**Mtb, *M. canettii and M. africanum***	***Wbbl1* gene**	**(Reddington, O'Grady et al. 2011)**
JOE probe	**JOE - ATATGAAA+CAG+CATGA+CTTT-BBQ**	**Inhibition control plasmid *pCR2.GFPRD4***	**Plasmid design region**	**(Pontiroli, Travis et al. 2011)**

+: indicates the insertion of an LNA base

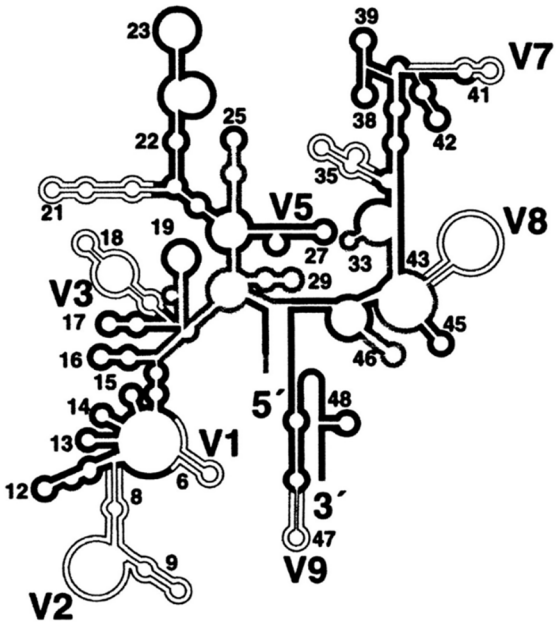


Figure 2.7 The secondary structure of Mycobacterium 16S rRNA and the helix 18 in V3 region highlighted in number 18 and V3 in bold was target region of APTK primer (Tortoli 2003).

2.5.2. qPCR approach and method

All DNA products using qPCR were loaded on to the MicroAmp 96-Well Reaction Plates (Applied Biosystems Inc., CA, USA) and sealed with the X100 Adhesive PCR Film Polyester (Scientific Laboratory Supplies Ltd, Nottingham, UK). 1 min centrifuge at

1250 RPM after plate setting-up and then placed into the ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA) for amplification of the DNA sequence. A $\log_{10}$ serial dilution of reference standard (10^6 to 1 genomic equivalents μl^{-1}) was measured using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Leicestershire, UK) with 2 µl sample loading. The DNA value (ng/µl) was estimated by measuring the absorbance at 260 nm UV wavelength. The yield of DNA was estimated using dsDNA copy number calculator (http://cels.uri.edu/gsc/cndna.html) to obtain the number of DNA copy of samples and then diluted to the reference standard as required.

All results for qPCR were analysed using the 7500 Fast System SDS Software (Applied Biosystems) and export to the excel 2013 file to perform further analysis and Figure. The Ct (cycle threshold) value was set up to 0.02ΔRn (logarithm of normalised reporter) to normalise the quantification during analysis. The fluorescence emission below the Ct and all negative template control (NTC) were be recognised as negative results. Comparison of data were performed using statistical analysis on pathogenic prevalence and environmental reservoir research.

Each sample was duplicated and run for positive-negative test at first with single Mtb reference standard (10^4 genomic equivalents μl^{-1}) for Mtb detection and BCG reference standard for Mb detection. Replicates was identified as potential positive result with at least one positive. These potential positive samples were triplicated and then test again on quantification using whole $\log_{10}$ dilution of Mtb reference standards for Mtb detection and BCG reference standards for Mb detection. Sample was identified as positive with all three technological replicates positive and the

mean and accumulation of genome equivalents was estimated for further statistical analysis.

2.5.3. Sensitivity and Specificity of qPCR assays

The primers and probes listed in Table 2.6 had proven their unique specificity on their targets in the previous studies, the RD9 on Mtb and *M. canettii* in Halse *et al*. (Halse, Escuyer et al. 2011) and RD4 scar on Mb (Sweeney, Courtenay et al. 2007). qPCR conditions were modified to increase the sensitivity. Primer and probe concentration, for instance, was optimised to 900 nM in the final reaction concentration on environmental sample screens.

The sensitivity test started with spikes Mtb and Mb cells seeded in liquid solution such as water and milk for sensitivity of different primers and probes. In brief, a $\log_{10}$ dilution of reference standards with each 200 µl volume from 10^7 to 1 cells μl^{-1} were seeded into each 1800 µl liquid sample, separately. A $\log_{10}$ dilution range was prepared from 10^6 to 0.1 cells μl^{-1} equal to 10^9 to 100 cells L^{-1}. DNA was extracted using FastDNA™ SPIN Kit for Soil and tested using qPCR with different primers and probes. These primers and probes consisted of the RD4 scar, RD9, LepA and Wbbl1 primers targeted on Mb, Mtb, Mb and Mtb, respectively and used to compare sensitivity. The qPCR programme and chemical reagents were as before (Travis, Gaze et al. 2011).

2.6. Internal control

JOE probe sequence (Table 2.6) was employed in this study targeted to design plasmid pCR2.GFPRD4 (Pontiroli, Travis et al. 2011)(Figure 2.8), with gene encoded region where RD4 scar primer attached and sequence area JOE probe targeted. The plasmid and JOE probe added as internal control to detect any inhibitors. The qPCR programme ran as follows, 50.0 °C for 2 min followed by 95.0 °C for 10 min then 50 cycles of 95.0 °C for 15 sec and 58.0 °C for 1 min until the end of programme with FAM reporter for fluorescence emission of RD4 scar probe and GFP reporter for fluorescence emission of JOE probe. Mixtures of qPCR reaction solution included 1 mg ml^{-1} BSA, 12.5 µl of TaqMan Environmental Master Mix 2, 10 µl log_{10} dilution of reference standard as before, 10 µl of plasmid pCR2.GFPRD4 of log_{10} dilution from 10^6 to 1 plasmid μl^{-1}, 900 nM of RD4 scar primers consisting of forward and reverse primer and 250 nM RD4 scar probe and 250 nM JOE probe made up to total 50 µl with sterilised molecular grade water.

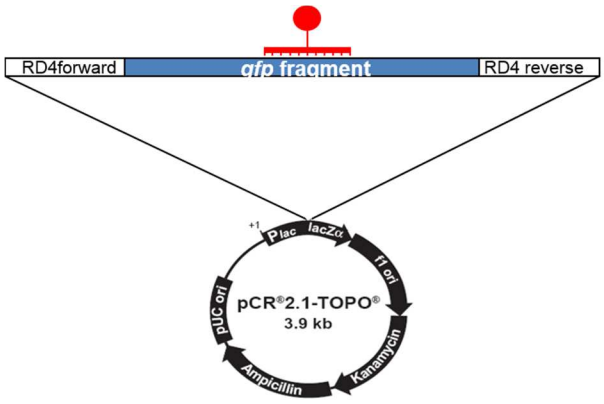


Figure 2.8 The designed plasmid pCR2.GFPRD4 with gene encoded region where RD4 scar primer attached and sequence area JOE probe targeted (Pontiroli, Travis et al. 2011)

2.7. Immunomagnetic capture (IMC)

A set of spikes for Mtb and Mb cells from 10^7 to 1 cells μl^{-1} were prepared and then seeded in liquid solution such as water and milk to investigate if it was possible to improve sensitivity with two antibodies, PAbs and MAbs. The first one was anti Mtb PAbs (PA1-7260) raised in rabbit against whole cell sonicates of Mtb (Thermo Fisher Scientific, Leicestershire, UK). The second was MAbs MBS 43 raised by the Veterinary Laboratories Agency against the MPB83, a glycosylated cell-wall associated protein known to be in the outer wall of Mb and Mb BCG. The PAbs were combined with Sheep anti Rabbit Dynabeads (Thermo Fisher Scientific, Leicestershire, UK) compared to the MAbs attached to magnetic beads using rabbit anti mouse Dynabeads (Thermo Fisher Scientific, Leicestershire, UK). The presence of antibodies in qPCR was compared to qPCR applied without antibodies. Each set of spikes were seeded into each 1800 µl liquid sample, separately and mixed and then adjusted to provide a dilutions series from 10^6 to 0.1 cells μl^{-1} equal to 10^9 to 100 cells L^{-1} and blocked with 500 µl 3 % BSA overnight at 4 °C. The dynabeads were mixed with PAbs overnight at 4 °C then washed with 200 µl PBS twice using Dynal MPC-L immunomagnetic capture device (Thermo Fisher Scientific, Leicestershire, UK) to ensure unattached antibody was removed. These beads were loaded into prepared bacterial solution and mixed, washed and rolled overnight until beads attached to the targets cells. The recovered beads were washed twice with PBS the next day and resuspended in 100 µl PBS and the DNA was extracted from beads using FastDNA™ SPIN Kit for Soil and tested using qPCR.

2.8. 454 pyrosequencing and QIIME analysis

2.8.1. Sample preparation for pyrosequencing

DNA from 58 samples was selected from 510 samples in the dry season. These samples consisted of 10 faecal samples, 12 household dust, 23 boma soil, 6 water and 7 sediment samples. The 10 faecal samples including 6 cattle and 4 goat faeces were selected from all faecal samples with the highest quantified positive detected by APTK primer which targeted SGM species. The 2 household dust and 4 boma soil samples were selected from each village for same reason. There are one water and sediment samples from each water sampling site. A total 15 µl of each sample was subjected to NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Leicestershire, UK) to achieve optimised quality and standardise concentration of DNA for pyrosequencing. These DNA samples were sent to MR DNA ((Molecular Research LP), Shallowater, USA) for pyrosequencing with specific primer APTK-pyro.

2.8.2. Pyrosequencing method

The DNA sequencing machine used by MR DNA was 454 Genome Sequencer Roche FLX machine with FLX Titanium system. The FLX Titanium system was capable of short running time < a day, and the DNA read length reached up to 1000 bp (700 bp in average) and 99.997 % accuracy.

2.8.3. QIIME bioinformation analysis

QIIME was an open-source bioinformation platform with large number of disparate programmes from different sources. This QIIME platform was used to input the raw 454 pyrosequencing data and different downstream analyses were used consisting of demultiplexing, quality filtering, OTU picking, taxonomic assignment, phylogenetic

reconstruction. The detail procedure of analysis and commands applied for QIIME (appendix 3 and Figure 2.9). In this study, QIIME and two other sequencing analysis software system systems were used; UPARSE and Oligotyping, to determine OTU groups, phylogeny compared to known species and complexes. Only Oligotyping was likely to be suitable for matching to species. The amplicon was specific for SGM.

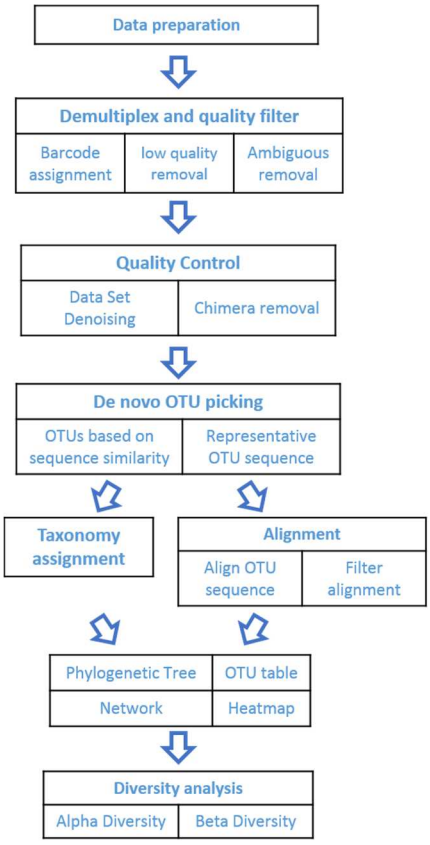


Figure 2.9 The detail procedure of QIIME on analysis of 454 pyrosequencing data.

2.8.4. QIIME: Preparation of raw pyrosequencing data

The Standard flowgram format files were obtained from 454 Genome Sequencer Roche FLX machine. The first step was to convert SFF file into fasta files and qualities file (qual) for the next step. Mapping files were retrieved from the 454 machine,

including all information on samples, such as barcodes, primer sequence and other information provided to analyse and cluster.

2.8.5. QIIME: Demultiplex and quality filter

The adaptor and barcode sequence were attached to the amplicon (Crawford, Crowley et al. 2009) (Figure 2.10) prior to 454 pyrosequencing. This approach was necessary for amplification and identification of each sample. Adaptors were removed after sequencing and barcodes excluded (Figure 2.10). The new barcode was introduced to each sequence for each sample and low quality and ambiguous reads removed using the Python script, *split_libraries.py*. The amplicon length in this study was between 420bp to 500bp according to the amplicon size without variation of insertion and deletion. The amplicon from raw data was chosen to be kept according to sequences matching the forward primer with a maximum of 1 mismatch to ensure the quality. These procedures was therefore kept in this step.

Figure 2.10 The example of amplicon construction consisted of the adapters, barcode sequence, primers and target sequence for 454 pyrosequencing (Crawford, Crowley et al. 2009)

2.8.6. QIIME: Quality Control

Accumulation of sequencing errors and chimeras produced during pyrosequencing amplification was unavoidable and lost accuracy. The aim of denoising and removal of chimeras was to eliminate the occurrence of inaccurate sequencing error. The Python script, *denoise_wrapper.py*, was applied for denoise using flowgram similarity

in order to exclude singletons in each data set. The script, *inflate_denoiser_output.py*, after denoise was to inflate the repeat numbers back to the data set to avoid being integrated during denoising step.

The Usearch61 was then employed on chimera sequence check and subsequently excluded chimera sequences before clustering sequence into each OTU. The Python script used in this step was *identify_chimeric_seqs.py* which checked against online chimera-free DNA and RNA reference sequence database from all organisms to perform chimeric sequence checking and removal (Edgar 2010). These steps were critical due to eliminate the proportion of random nucleotide insertion and deletion occurred in the amplicon sequence.

2.8.7. QIIME: De novo OTU picking

All sequence reads from diverse samples were clustered together into OTU based on their sequence similarity and variation. The python script in OTU picking was *pick_otus.py* with parameter, 0.97 referred to those sequence with 97% similarity were clustered together into one OTU group. This OTU picking programme, Uclust, was reference-based method with reference from the Greengenes database (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi). A representative sequence was selected based on most abundant sequence in each OTU for downstream analysis using the python script, *pick_rep_set.py*.

2.8.8. QIIME: Taxonomy assignment

The identification of each representative sequence to the taxonomic group was done to convert and classify each OTU to known organism or uncultured groups based on different databases (Table 2.7). The BLAST taxonomy assignment was applied, and

also uclust consensus taxonomy assigner (default) and the RDP classifier. The BLAST assignment was against the online reference, BLAST database using the python script, *assign_taxonomy.py*, with parameters, *-m blast*, to change default method to the BLAST assignment.

Table 2.7 Different taxonomic assignment methods based on their unique database and threshold to classify the OTU to known organism. The time period of process was estimated by Intel® Core™ i5-3337U CPU@ 1.8GHz prosessor.

	Database	**Threshold**	**Time period**
Uclust	Greengenes, RDP and SILVA	Similarity & hit	Middle (~1hr)
BLAST	NCBI	E-value	Slow (~3 hr)
RDP	RDP	Confidence score	Fast (~30 min)

2.8.9. QIIME: Alignment, filter and Phylogenetic Tree

The default PyNAST alignment python method was implemented with Nearest Alignment Space Termination alignment algorithm (Caporaso, Kuczynski et al. 2010) with python script, *align_seqs.py*. The reference in this alignment was "core_set_aligned.fasta.imputed" from Greengenes database which contained > 200,000 non-chimeric candidate sequences. The alignment was completed via insertion of gaps to the template sequence and used for downstream analysis such as diversity analysis and Oligotyping. Every sequence after alignment was filtered and the redundant gaps removed so the typical sequence varied between 200-400 bp depending analysis tool. The landmark file was supplied and defined which position being reserved while building the phylogenetic tree.

The phylogenetic tree based on OTU produced using default FastTree with the python script, *make_phylogeny.py*.

2.8.10. QIIME: Alpha and Beta diversity analysis

Several steps were included in alpha diversity analysis such as generate rarefied OTU table, compute measures of alpha diversity and generate alpha rarefaction plots. The algorithms for alpha diversity analysis in order to compute data matrics were shannon, PD_whole_tree, chao1 and observed_otus. The algorithm selected in this study was Shannon diversity indices because the application of this algorithm served the comparison of both species richness and abundance from diverse habitats or samples sets.

The PCoA three dimension plots were represented for beta diversity results. These results of beta diversity included two different phylogenetic matrics; Weighted UniFrac and Unweighted UniFrac measure. The Weighted UniFrac measure looked at the relative prevalence of OTU while the Unweighted analysis is a qualitative test concerned about the presence/absence of OTU.

2.9. Oligotyping

The clustering method from OTU was based on taxonomic assignment and sequence similarity. This assignment method provided poorly resolved diversity description (Eren, Maignien et al. 2013) and represented only a small proportion of the estimated microorganism diversity in environmental sample screening (Huse, Welch et al. 2010). In addition, the sample screening based on 16S rRNA used in this study was of limited

specificity because of identical 16S rRNA sequences within the members of MTBC but high sensitivity with single nucleotide difference at 16S rRNA level (Thompson, Pacocha et al. 2005). For this reason, the clustering approach based on single variable site in each sequence read via entropy analysis was appropriated for environmental sample screening to resolve species diversity.

2.9.1. Oligotyping: Alignment and trimming

The Oligotyping pipeline was very flexible and was able to start after the denoising step from QIIME platform or samples pooling step from UPARSE platforms. The file format used in Oligotyping was single Fasta format file generated from both platforms. Only one step was required to change from UPARSE. The position of the sample name from the default position which following by system label name generated automatically by MiSeq and changed to be prior to system label name using the LINUX terminal command, *sed 's* during UPARSE demultiplexing step. It was because the identification of Oligotyping approach only recognised the sample name in the particular position.

Alignment of all amplicon reads was necessary and critical during entropy analysis in Oligotyping. The bases in each read were required to be in the correct position without interference of random insertion/deletion errors for entropy analysis to be applied. The reference, "gg_97_otus_6oct2010_aligned.fasta" with > 500,000 aligned sequence from Greengenes database was compared to "core_set_aligned.fasta.imputed" from same database used for alignment in this study to estimate which one is more efficient on alignment. The sequence trimming subsequently was performed to trim each amplicon reads into same length with

uninformative gaps insertion or deletion using Oligotyping script, *o-trim-uninformative-columns-from-alignment* before entropy analysis.

2.9.2. Oligotyping: Shannon Entropy Analysis

The one most important part of Oligotyping was Shannon entropy analysis. This analysis has introduced a new dimension for sequence analysis at the sub-genus level. The approach focuses on single nucleotide position of interest and accumulation of positional variation to perform an algorithm to identify highly refined taxonomic units called oligotypes. This entropy analysis was distinguished from the clustering approach which is based on sequence similarity and taxonomic assignment used in QIIME and UPARSE platforms. The Shannon entropy analysis is based on equation 2.2 described below:

$$H = -\sum_{i} p_i \log_b p_i$$

Equation 2.2.

The Shannon entropy (*H*) was given by this formula and *Pi* is the probability of positional variation within the one sample set. Application of summation (Σ) in this equation helped the accumulation of positional variation and distinguished it from other random errors occurred under the large data set.

The oligotypes script in this step was simple to run and entitle, *entropy-analysis*, and at least five highest positional variations were selected to perform downstream Oligotyping.

2.9.3. Approach and methods

The Shannon entropy analysis was used to eliminate as most reads containing random sequence errors (when 454 pyrosequencing was used). However it was

difficult to provide error-free data set when the completed discrimination of the sequence errors had been completed. The selection of the positional variation was critical and the parameters used for Oligotyping. The analysis step was used on environmental diversity analysis by running the script, *oligotypes*, with the parameters. The parameters included in to Oligotyping were functional and diverse for different purposes, for instance, *M* indicated a limit for the least abundant oligotypes to reduce the chance noise had occurred and *C*, helped to select positional variation of interest for the oligotypes establishment. The online database BLAST was applied in default for organism assignment. The oligotypes were identified as 100% sequence identity and 100% query coverage to each species based on this online database BLAST. The proportion of different organisms in each sample was shown with static HTML page if *–gen-html,* this parameter was included in the command line. The similarity of each sample was based on the proportion of organisms obtained from Oligotyping analysis. The results were collected, clustered and compared to establish the clustering tree. This tree provided the evidence to estimate how each sample was similar or different on diversity analysis.

2.10. R: the project for statistical analysis and graphical presentation

R is a programme language and software environment for statistical computing and graphics. The application of R was capable of statistical techniques including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification and cluster. The R programmes was also conducted functional spatial study based on different functional package (library) and algorithm choices. The libraries were

selected for spatial analysis in this study was *ggplot2*, *ggmap*, *raster* and *RColorBrewer* for diverse purposes. The *ggplot2* library was applied for data presentation, graphic model and map plotting. The *ggmap* package was delineated the spatial analysis on combination between google map and coordinate dataset. The R programme only recognised particular format so it was very important to convert dataset to appropriate format before importing files into R programme. The GPS coordinate, for instance, was recoded and formatted using Garmin eTrex® 10 handheld GPS device while samples collection and required to converted these GPS coordinates to decimal degrees including latitude and longitude before importing data into R programme.

The R script was text-based format helped to delivered multiple and continuous commands in one functional application including dataset input, convert and analysis. The appropriate libraries were selected for different purposes using the R script because the commands were variable depending on different libraries applied in the script. The incorrect library caused function corrupted while script delivered into the R. The R programme helped to establish the correlation between pathogens and environmental reservoir, for instance, the prevalence of pathogens was illustrating and highlighted in colour and size with google map using R programme. All details of R scripts created in this study are listed in appendix.

2.11. Additional statistical analysis

The quantification of Mtb and Mb genome equivalents was calculated to represent cells in each comparison for different tests. Differences in quantitative recovery of

DNA yields was analysed using the student t-test, the ANOVA and the paired t-test if the sample population was normal distribution. The normality test in statistics was used to determine whether the sample population followed a normal distribution or not. Another set of non-parametric statistical analyses used on non-normal distribution sample sets were the Mann Whitney (U) test, the McNemar's test and the Wilcoxon signed-ranks test. All statistical analyses were performed using OriginLab (Silverdale Scientific Ltd., Bucks, UK). The number of asterisks represented the different level of significant difference using statistical analysis such as an asterisk (*) if p value < 0.05, double (**) if p value < 0.01 and triple (***) if p value < 0.001.

Chapter 3 Optimisation of molecular approaches to detect *Mycobacterium* species in the environment

3.1 Introduction

The challenge for environmental pathogen reservoirs requires the rapid, reliable and accurate assessment of prevalence in diverse environmental sample types. It was accepted that the traditional liquid and solid culture applied to environmental samples was not feasible for the reasons given above. In addition it is well proven that cultivation approaches are extremely limited for diversity studies, particular in soil around 10^6 bacterial strains isolated using specific media or broth from cell cultivation and only around small proportion (1%) are culturable (Torsvik, Goksoyr et al. 1990, L.R. Bakken 1997). In addition, this cell cultivation based approach and isolation was employed rapid growing bacteria would dominate their more slow-growing counterparts.

The target pathogens in this study were two SGM, Mtb and Mb cell. These two SGM human and animal lethal pathogens were problematic and no adequate selective media to recover these cells from environmental samples due to their long doubling time around of 24 hours and the appearance of a single colony on media taking two weeks to a month. Another challenge for isolation of SGM species was high contamination rate from environmental samples on selective media, especially from faecal samples because Mb become sensitive to the usual decontamination methods using clinical isolation techniques (Young, Gormley et al. 2005). Decontamination is necessary for clearance of competing rapidly growing bacteria. The decontamination of samples consisted of the addition of 5 % NaOH following by H_2SO_4 and quaternary

ammonium compounds (Young, Gormley et al. 2005). Therefore, isolation of SGM cells is problematic and time-consuming because of growing-time, and few selective media exist such as Middlebrook media to recover slower growing mycobacteria cells.

The conventional cultivation techniques were not able to provide accurate estimates of species richness and evenness. Solutions to this issue are time-consuming and widely discussed. One remedy is to develop reliable, rapid and accurate molecular techniques. Another advantage is that animal environmental shedding can be used as proxy for tracheal shedding (Hayley King 2015) and samples examined using molecular detection prevent unnecessary invasive detection and diagnosis (trapping and blood testing).

Monitoring and prevalence analysis was done using samples of dust and body fluids (where available). Environmental monitoring was conducted through DNA isolation and targeted qPCR (Hayley King 2015). However the procedure also provided its own challenge due to different manufactures for individual manual DNA isolation method. The selection of kits was necessary to ensure DNA isolation was transferable to other labs and gave high quality DNA. Another task was to avoid PCR inhibitory compounds which existed in liquid and solid samples, such as humic acid in soil (Wilson 1997).

There are currently several methods for direct DNA extraction from liquid samples such as water and milk (Quigley, O'Sullivan et al. 2012). Methods for solid samples such as faeces, dust, soil and sediment were previously investigated (Neill 2010, Travis, Gaze et al. 2011). One is mechanical lysis approach and another is chemical based approach as described below.

1. Mechanical and physical lysis procedure using Ribolyser with tiny glass, ceramic or steel breads mixed with samples to achieve DNA isolation. This DNA isolation approach can release intercellular components via beads colliding with cells.

2. Chemical lysis approaches using lysis buffer normally includes detergent, sodium, chloride, EDTA and enzyme to degrade protein and RNA. The detergent works with EDTA to destroy microbial lipid bilayers of cell membranes and inhibit DNase activity via capturing divalent cations such as Mg^{2+}. And the enzymes, proteinase K, normally is applied to remove cellular and nuclear protein to prevent this material attaching to DNA and interfering with isolation. The ribonuclease RNase A can hydrolyse RNA molecules into single nucleotides to remove RNA contamination.

A comparison was performed to select which approach and kit could provide high yields of clean DNA of around 10 µg g^{-1} soil. However the differentiation of Mtb and Mb cell from other MTBC members presents a challenge due to their identical 16S rRNA. Four published primers and probes targeting different regions of chromosomal deletion and selected functional genes have been proposed for differentiation of MTBC include RD4 scar (Sweeney, Courtenay et al. 2007), RD9 (Halse, Escuyer et al. 2011), LepA and Wbbl1 (Reddington, O'Grady et al. 2011). All can be used for end point and qPCR. In the current study detection limits were required for use of these primers with DNA extracted from soils, faeces and sediments. The aim of the analysis was to select a detection systems for each target, Mtb and Mb for use in environmental monitoring.

The specificity of primers RD9, RD4 scar, LepA and Wbbl1 was reported (Sweeney, Courtenay et al. 2007, Halse, Escuyer et al. 2011, Reddington, O'Grady et al. 2011) and the detail of sequence and target pathogen species (Table 2.6). The RD4 scar primer was developed for specific detection of Mb cells from faecal samples of badgers (*Meles meles*) in the UK (Sweeney, Courtenay et al. 2007). The same RD4 scar primer was used for detection of Mb from soil and water filter samples in Ethiopia (Khera 2012) and for badger BCG vaccine trial in the UK (Hayley King 2015) and showed specificity in all cases.

The use of the RD9 primer was reported by Halse *et al* (Halse, Escuyer et al. 2011) for detection of Mtb cells in clinical specimens sample consisting of 94.3% respiratory specimens and 5.7% other types, such as abscess, aspirate, lymph node, body fluid, gastric fluid, urine, and wound specimens. The specificity of this primer was 97% and it successfully amplified Mtb in 372 samples

The LepA and Wbbl1 primers and probes were developed for detection MTBC and Mtb genome (Reddington, O'Grady et al. 2011). The Wbbl1 assay was specific for the detection of 26 Mtb strains and 5 *M. canettii* strains and no detection of other non-MTBC pathogens known as non-tuberculosis mycobacteria which can cause similar disease to Mtb. Another primer and probe, LepA, was applied for specific detection of MTBC member and all 64 MTBC strains were differentiated from 61 closely related NTM strains using LepA.

Another challenge was to avoid PCR inhibitory compounds which existed in liquid and solid samples to cause false negative on qPCR approach. These inhibitors even in very low concentrations could cause false negative on qPCR analysis. A remedy was

provided in this study as the specific antibodies based approach, IMC was performed to capture, concentrate and purify target bacterial cell prior to DNA extraction. Another approach to reduce false negative results was an internal control plasmid incorporated in the same PCR reaction tube with the sample. An internal control based on exotic gene fragment from green fluorescent protein flanked by RD4 scar priming sites was used previously but not in the same reaction mix (Pontiroli, Travis et al. 2011). Therefore, experiments for detection of Mb were done to improve the RD4 scar method by incorporating internal control plasmid in with the sample and multiplex the two, saving time and money. In order to optimise these experiments were done to determine if the internal control as target for the same primers competed with the sample target and reduced sensitivity.

3.2 Aims

1. To test different DNA isolation kits on the liquid and solid samples to reveal the most efficient DNA extraction kits for environmental screening

2. To examine the sensitivity and specificity of each primer and probe to target Mtb, Mb and other MTBC members in environment screening

3. To validate sensitivity using immune-magnetic beads application prior to qPCR detection in liquid samples

4. To prevent false negative via internal control added with qPCR approach

3.3 Results

3.3.1 Comparison of different DNA isolation kits in diverse environmental samples

Both mechanical and chemical lysis procedures are used in the FastDNA SPIN kit compared to only chemical lysis approach applied in the Blood & Tissue kit. The Mtb cells were resuscitated in liquid Middlebrook 7H9 broth with 10 % Middlebrook ADC enrichment for four weeks and the two different commercial kits were compared using qPCR approach with the specific primer RD9 for detection of Mtb genome equivalents (Figure 3.1). The result illustrates approximately three fold higher gene copies ml^{-1} using the FastDNA SPIN kit compared to the Blood & Tissue kit.

The analysis shows the FastDNA SPIN kit is more efficient than Blood & Tissue kit on DNA isolation from liquid culture. This result also implies application of both DNA lysis procedures consisting of mechanical and chemical lysis procedure is more effective than only chemical approach employment.

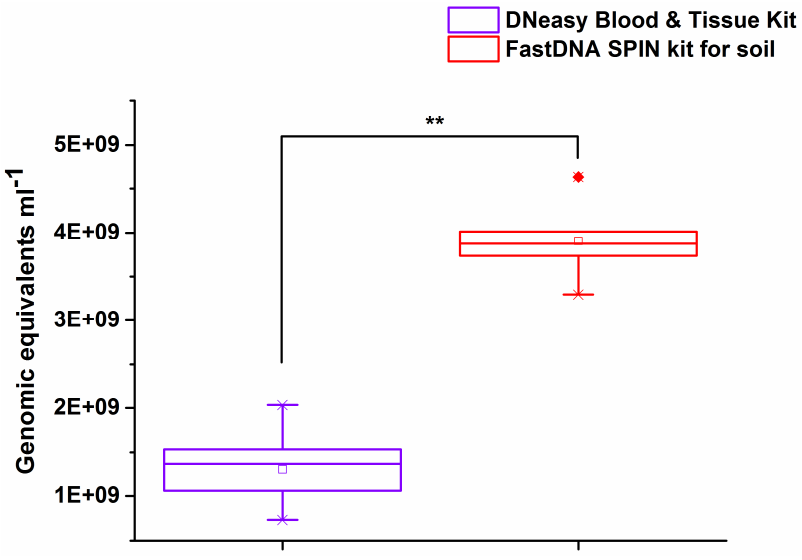


Figure 3.1 Comparison of both DNA isolation kits for quantification of Mtb genome equivalents. **Indicates significant difference (P-value < 0.01) using Mann Whitney statistical analysis.

Another liquid environmental sample type is milk which is aqueous fluid that contains dissolved carbohydrates and protein aggregates with minerals (Jost 2007). The DNA isolation from milk was difficult due to the frequency of occurrence of PCR inhibitors such as proteins, fats and minerals, calcium can compromise the efficiency of qPCR (Wilson 1997). There are diverse DNA isolation approaches including commercial kits and manual methods using phenol-chloroform and ethanol purification. Three different DNA isolation kits were used and compared in this study and included Powerfood® DNA Isolation Kit, Blood & Tissue kit, manual method from Cremonesi *et al* (Cremonesi, Castiglioni et al. 2006) and for the method O'Neill (Neill 2010) (Figure 3.2).

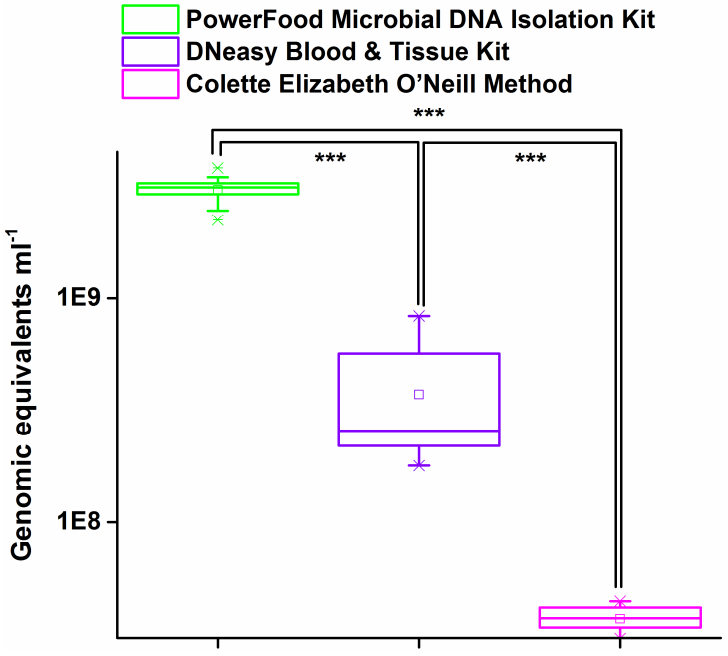


Figure 3.2 Comparison of three DNA isolation kits for quantification of Mtb genome equivalents from milk spikes. ***Indicates significant difference (P-value < 0.001) using Mann Whitney statistical analysis.

The Mtb cells were seeded into the raw milk (Walton Lodge Farm, North Crawley, UK). The DNA isolation kits were compared using qPCR approach with the specific primer RD9 for detection of Mtb genome equivalents with three biological replicates.

The yield of DNA genome was variable with each method and the PowerFood kit was more successful than other two kits at $3.04x10^9$ genome copies ml^{-1} on average. The O'Neill Method in contrast provided the lowest yield of DNA genome at $3.65x10^7$ genome copies ml^{-1}. The Blood & Tissue kit was tenfold better than the O'Neill Method at $3.71x10^8$ genome copies ml^{-1}. Significant difference between each kit (P-value *< 0.001*) marks with triple asterisks.

The DNA isolation kit used for solid environmental samples such as faeces, soil, household dust and sediment was FastDNA SPIN kit for soil. The use of this kit for DNA isolation from faecal samples was reported by Pontiroli *et al* (Pontiroli, Travis et al. 2011)and further improved by Travis *et al* who reported good sensitivity and efficiency of extraction for this kit (Travis, Gaze et al. 2011).

The PowerWater® DNA Isolation Kit was utilised on DNA isolation and extraction from triplicate water filter membranes from each water sampling location. The efficiency and sensitivity of DNA extraction was well proven by Khera (Khera 2012). A comparison between PowerWater DNA Isolation Kit and the other two manual methods the adapted Griffiths (Griffiths, Whiteley et al. 2000) and Pickup (Pickup 2004) methods was reported. The results showed that PowerWater kit was the most sensitive and effective method for detection of Mb genome at 10^2 to 10^5 Mb gene copies g^{-1}, compared to the Griffiths (10^4 to 10^5 gene copies g^{-1}) and Pickup methods (only 10^5 gene copies g^{-1}).

The different DNA isolation kits for diverse samples were therefore FastDNA for solid samples, PowerFood DNA for milk samples and PowerWater DNA for filter membranes from water samples.

3.3.2. qPCR Sensitivity of SGM species and MTBC

3.3.2.1 qPCR sensitivity from sterilised water

The RD9 primers were only applied in clinical samples and there were no reports of prior use for environmental screening. This study is first reported application of these primers in environmental samples. It is crucial to understand the limit of detection for each primer on target bacteria from environmental reservoirs to assess sensitivity and specificity. The sensitivity of each primer was estimated for the lowest recovery rate from $\log_{10}$ serial dilution range from 10^7 to 10^3 spiked cells ml^{-1} in liquid and 10^8 to 10^3 spiked cells g^{-1} in solid samples.

The triplicate of Mtb and Mb spiked water was subject to DNA extraction by FastDNA and the sensitivity determined using each specific primer and probe. The RD9, RD4 scar, LepA and Wbbl1 were examined, and then compared (Figure 3.3 and 3.4).

The minimum concentration detected by these four primers RD4 scar, RD9, LepA and Wbbl1 were $2.17x10^3$, $6.74x10^3$, $1.18x10^4$ and $6.30x10^4$ genome equivalents ml^{-1}, respectively (Figure 3.3) with no detection of spikes less than $2.17x10^3$ gene copies per ml^{-1}. This Figure 3.3 shows the highest sensitivity within these four primer is RD4 scar, more efficient than LepA for detection of Mb cells. The RD9 primer for detection of Mtb genomes is more effective than the Wbbl1. The efficiency of each primer was depicted in Figure 3.4 with the proportion of detection on spiked samples. The

sensitivity helped to select the most effective primer for environmental samples screening.

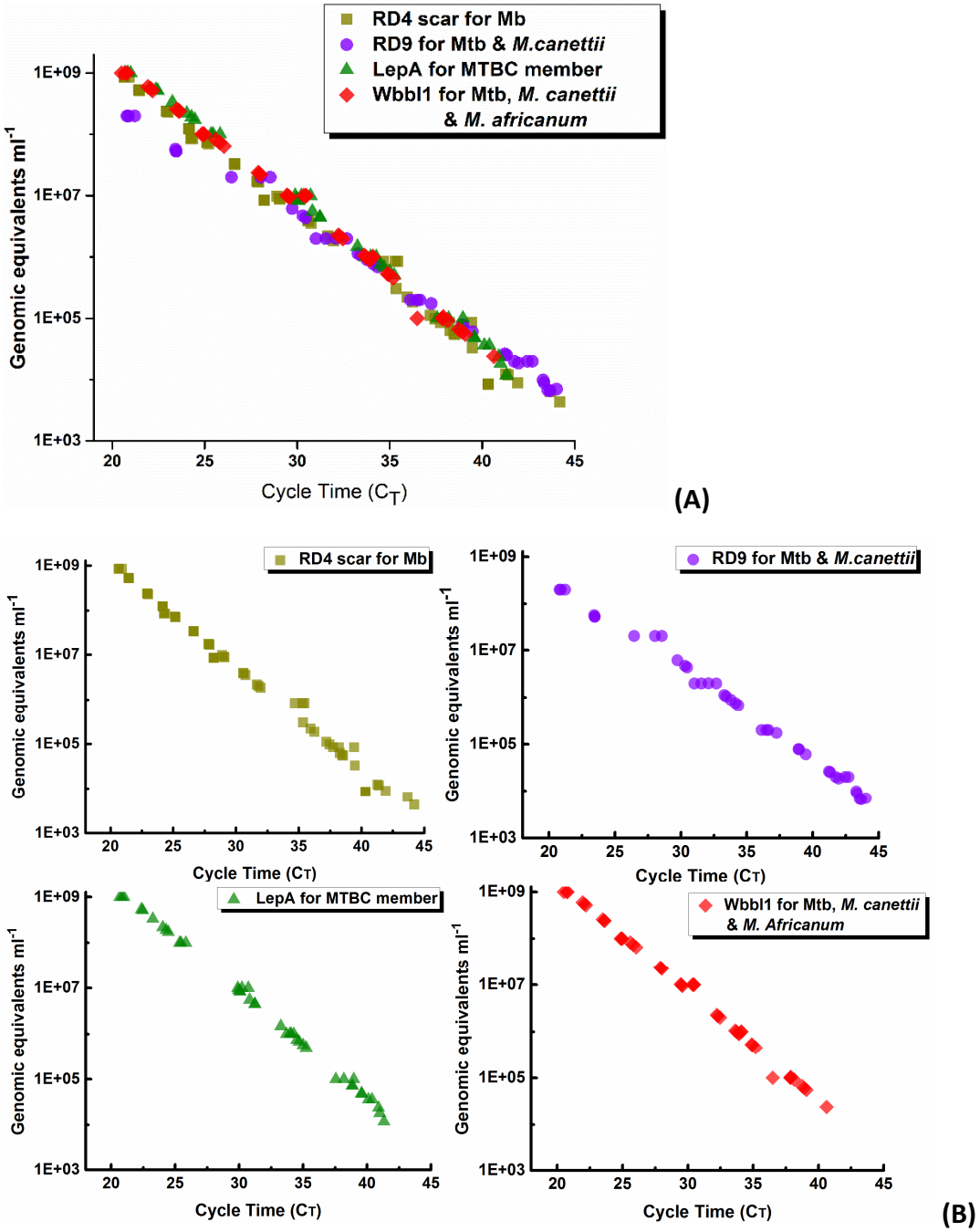


Figure 3.3 Investigation of sensitivity of RD4 scar, RD9, LepA and Wbbl1 for detection of Mb and Mtb cells in sterilised water samples (A). The Figures are shown separately in (B) for clarity.

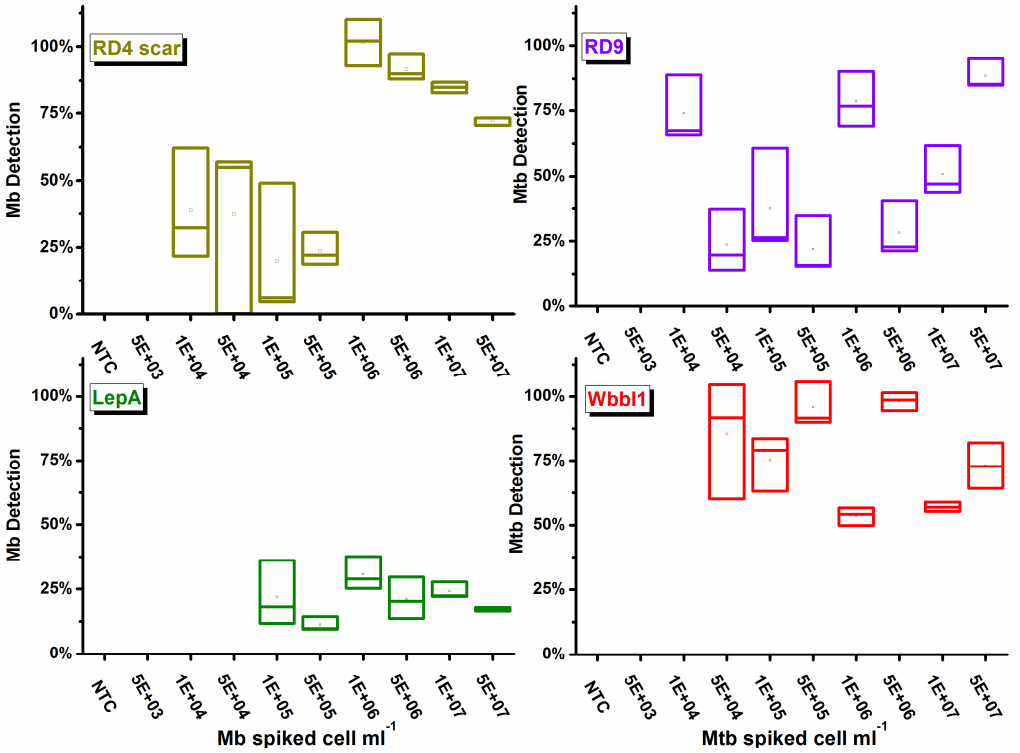


Figure 3.4 Detection of Mb and Mtb cells from 1.0 ml of spiked sterile water at a range of cell counts ml^{-1} using qPCR with different primers and probes.

The range of Mb spikes detected by RD4 scar was from $5x10^3$ to $5x10^7$ gene copies ml^{-1} compared to LepA in the same condition for detection of same Mb spikes. The $1x10^6$ group can reach 100% recovery in terms of RD4 scar detection compared to LepA with all groups less than 40 %. The best recovery was $1x10^6$ copies ml^{-1} for detection of Mb with RD4 scar primer.

The RD9 detection of Mtb gene copies was more sensitive than Wbbl1 primer but the latter performed at high cell counts.

The RD4 scar and RD9 primers were considered as optimal for use with liquid environmental samples. A further experiment was applied to seed Mb and Mtb spikes in raw milk which contained PCR inhibitors to explore the limitation and efficiency of detection with the RD4 scar and the RD9 primer.

3.3.2.2. qPCR sensitivity from raw milk

The PowerFood kit was employed for DNA extraction from unsterilized milk products after Mb and Mtb spikes seeded in raw milk. The specific primers RD4 scar and RD9 selected from previous trials with water were applied for detection in milk and tested for sensitivity (Figure 3.5) and recovery rate (Figure 3.6).

The sensitivity of two primers was tenfold less in unsterilised milk compared to sterilised water, for instance, the minimum concentration of Mb detection by the RD4 scar was $6.77x10^4$ gene copies ml^{-1} in milk compared to $8.50x10^3$ copies ml^{-1} in water. The same situation occurred for detection of Mtb genome with the RD9. The limit of detection of RD9 was $8.59x10^4$ copies ml^{-1} in milk compared to $9.9x10^3$ copies ml^{-1} in water. It might be suggested that the incomplete lysis and interference of PCR by inhibitors contained in raw milk.

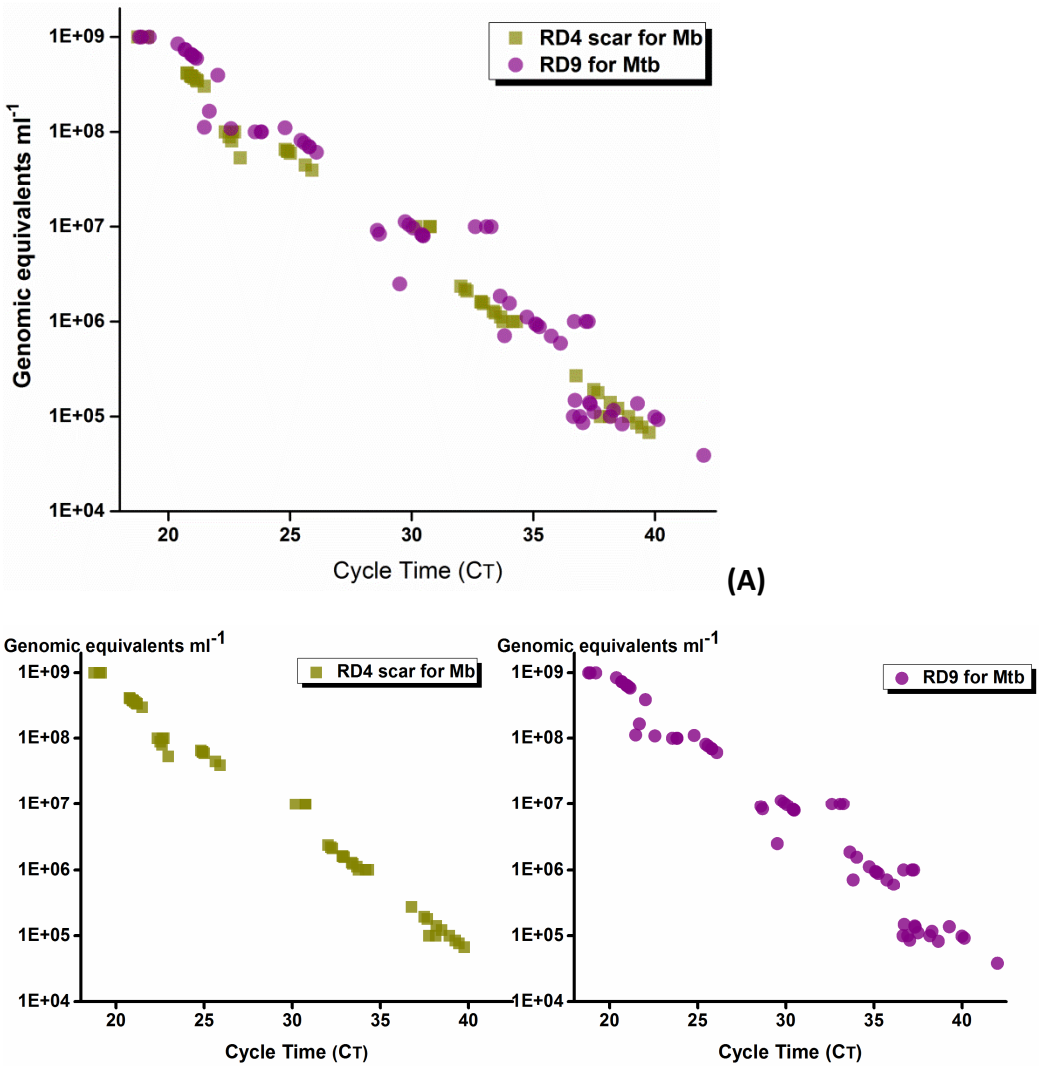


Figure 3.5 Investigation of sensitivity of RD4 scar and RD9 for detection of Mb and Mtb cells in raw milk samples (A). The Figures are shown separately in (B) for clarity.

The detection studies revealed in Figure 3.5 indicated that there was almost 0% detection below $1x10^5$ for RD4 scar. The test performed better at higher population levels.

The RD9 primers performed better and reached 100% detection or above due to bTB infection in the cattle unknown at the term of sampling. There was no detection from qPCR non-template control so the occurrence of contamination was excluded. The

$6.88x10^4 \pm 2.12x10^4$ Mtb gene copies ml^{-1} detects from raw milk without Mtb cells seeded.

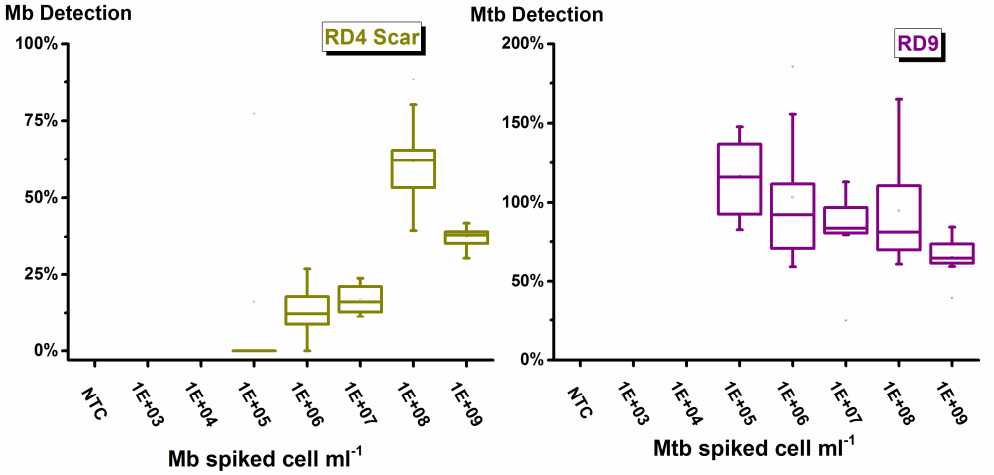


Figure 3.6 Detection of Mb and Mtb cells in raw milk at a range of cell counts ml^{-1} using qPCR with two different primers and probes.

3.3.2.3. qPCR sensitivity from soil

The next step aimed to examine sensitivity and recovery rate from solid samples with two specific primers. A previous study provided the evidence for detection of Mb genomic equivalents from faecal samples using qPCR with RD4 scar primer (Travis, Gaze et al. 2011). The experiment was conducted to validate the RD9 primers for detection of Mtb. A log10 serial dilution of Mtb cells was seeded into sterilised Warwick soil from Cryfield site, University of Warwick, UK (Pontiroli, Travis et al. 2011). The concentration was adapted to 10^4 to 10^8 cells g^{-1} and the DNA was extracted using FastDNA kit and tested using qPCR with the RD9. The sensitivity of the RD9 is illustrated in Figure 3.7 and 3.8.

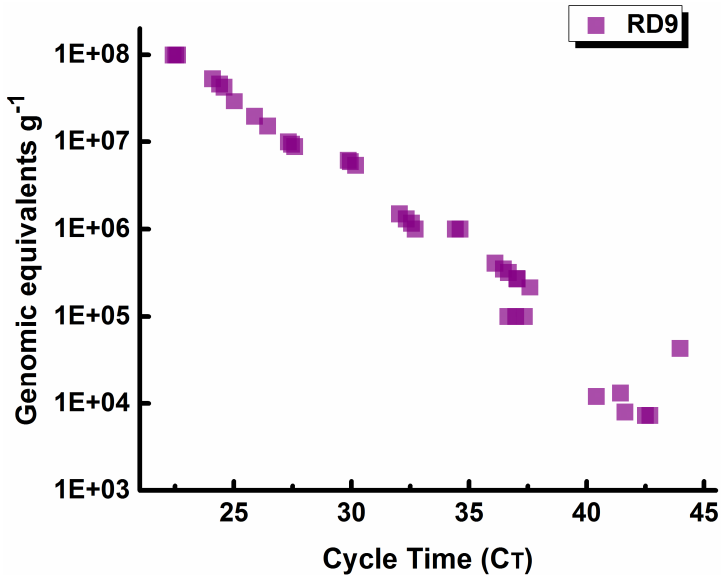


Figure 3.7 Investigation of sensitivity of RD9 for detection of Mtb cells in Warwick soil samples.

The sensitivity based on minimum genomic titre of Mtb recovered by the RD9 was $7.32x10^3$ gene copies per g^{-1} in soil and it was similar but more sensitive in water at $9.9x10^3$ copies per ml^{-1}. The previous study indicated there was no detection of sample with spikes lower than 10^5 cells g^{-1} with RD4 scar (Travis, Gaze et al. 2011). The RD9 detection was below 10^4 cells g^{-1} and this was a result of further optimisation compared to previous work (Halse, Escuyer et al. 2011).

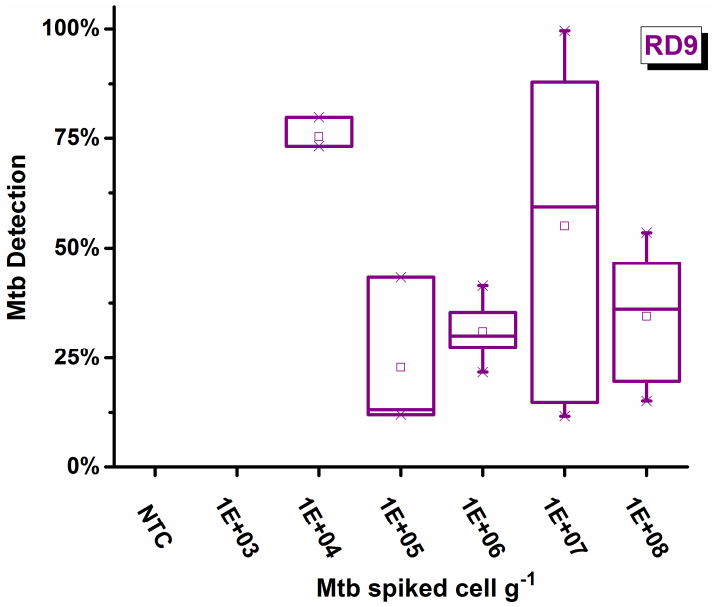


Figure 3.8 Detection of Mtb in Warwick soil using qPCR with RD9 primers and probes.

Most of the RD9 sensitivity was still lower than 50% detection but the positives of 75% at 10^4 copies g^{-1} was similar to the sterilised water test at 80% recovery.

The RD4 scar and RD9 revealed high specificity and sensitivity for Mb and Mtb detection. The minimum concentration detected was below 10^4 gene copies ml^{-1} in liquid and 10^4 gene copies g^{-1} in solid samples. The low sensitivity and recovery rate in unsterilised milk; however, must be addressed using different approaches. For this reason, IMC, the antibody based technique was applied in raw milk to test whether it can improve the sensitivity and reduce inhibition of PCR.

3.3.3. qPCR and IMC specificity and sensitivity: spiking cells in environmental samples

The sensitivity of these specific primers was provided in qPCR approach in water, milk and soil. A previous study indicated the IMC technique was successfully adapted to isolate Mb cells from naturally infected soil and badger faeces (Sweeney, Courtenay et al. 2006). But it is still uncertain if combination with antibody capture based method can validate sensitivity of qPCR in liquid samples. The IMC was antibody based technique using specific antibodies to concentrate target cells in liquid and then were collected with magnetic beads. One of the advantages for IMC-qPCR was to avoid contamination with other bacterial cells prior to DNA extraction.

3.3.3.1. IMC: The sensitivity and specificity of each antibodies for detection of Mtb and Mb

Two antibodies including MAbs raised against Mb and PAbs raised against Mtb were used for concentration of Mtb and Mb cells before qPCR method.

There were two approaches for indirect capture with antibodies used in this IMC, the first group was PAbs and MAbs for concentration of target cells. The second group of antibodies designed to capture first antibodies and attached to magnetic Dynabeads. The IMC aims to examine which specific antibody is more efficient for concentration of *Mycobacterium* species in varying ratios of mixtures of Mtb and Mb in sterilised water (Figure 3.9).

The sensitive and specific antibody was compared between IMC-qPCR and traditional qPCR for detection of bacterial cells in spiked raw milk (Figure 3.10).

Both Mtb and Mb cells proliferated for four weeks were seeded into sterilised water and were mixed with PAbs and MAbs before DNA extraction. The RD9 and RD4 scar qPCR were applied for detection of Mtb and Mb respectively with and without IMC. The recovery rate between PAbs and MAbs was compared (Figure 3.9). Higher cell detection occurred in the IMC using the PAbs compared to the Mabs; cell recovery with RD9 using PAbs was 75% while only 1% using MAbs. In addition, the PAbs raised in Mtb were capable of capture for both Mtb and Mb cells because multiple epitopes were recognised by the PAbs. This character made PAbs more flexible on capture of target cells.

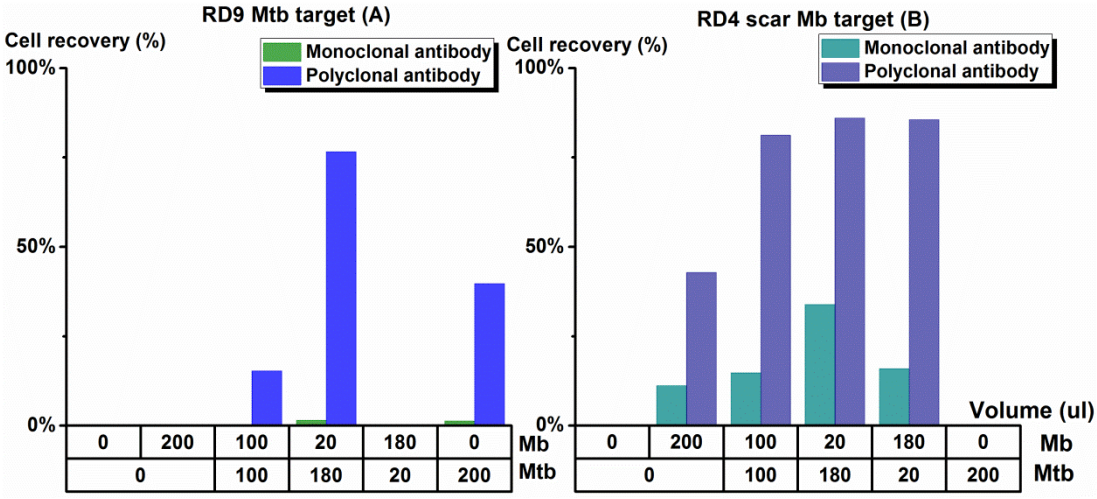


Figure 3.9 Comparison of efficiency and selectivity of MAbs and PAbs in sterilised water using the RD9 and RD4 scar qPCR for enumeration.

3.3.3.2. IMC: The sensitivity of IMC-qPCR and qPCR compared for detection of Mb cells

RD4 scar qPCR applied for detection of Mb cells in raw milk was problematic due to the high Mtb cell counts present in raw milk which caused the competition between

original Mtb and seeded Mtb spikes. This competition resulted in inaccurate comparisons of PAbs and qPCR for detection. The comparison of sensitivity between IMC-qPCR and qPCR approach (Figure 3.10) shows the limit of detection by IMC-qPCR was $3.42x10^3$ gene copies ml^{-1} compared to $1.20x10^4$ gene copies ml^{-1} using qPCR. There was fourfold higher sensitivity with the IMC-qPCR than in the qPCR approach.

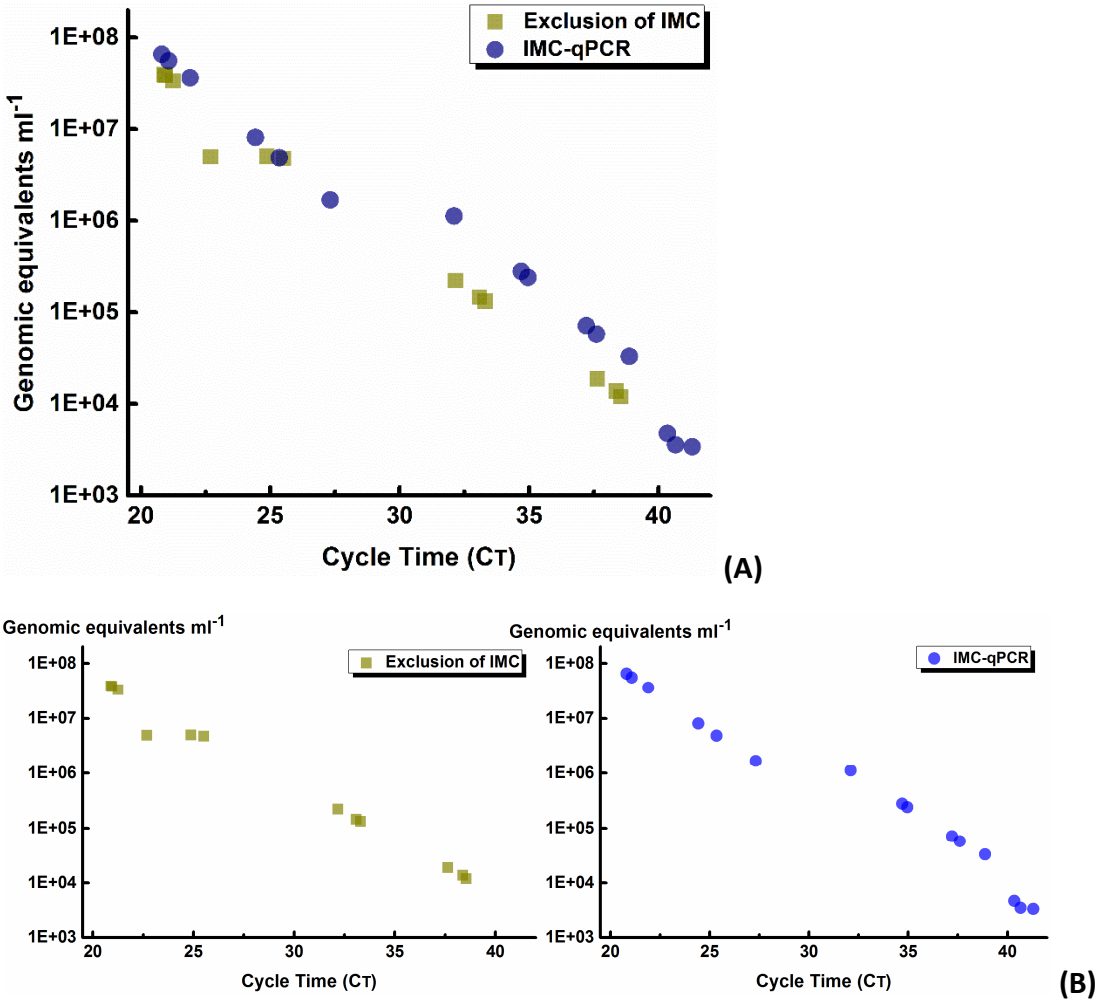


Figure 3.10 Investigation of sensitivity of RD4 scar qPCR with and without IMC for Mb cells detection **(A)**. The Figures are shown separately in **(B)** for clarity.

The detection demonstrates that IMC successfully concentrates Mb cells and the sensitivity was enhanced from 10^5 to 10^4 genomic equivalents per ml^{-1} with unsterilised milk (Figure 3.11). In addition higher cell detection was observed at 50%

on average from 10^4 to 10^8 cells ml^{-1} with PAbs capture. In contrast to IMC-qPCR, most of cell capture using the qPCR approach achieved below 50% capture with only 20% capture on average at 10^5 cells ml^{-1}. It is certain that there was more than 40% cell detection in 10^4 cells ml^{-1} spikes in raw milk on IMC-qPCR with no detection using the traditional approach. These results demonstrate that IMC enhances sensitivity and concentrates target cells. The IMC was required in certain liquid samples in this trial for example raw milk but for majority of samples it was not worth the cost, especially for solid samples.

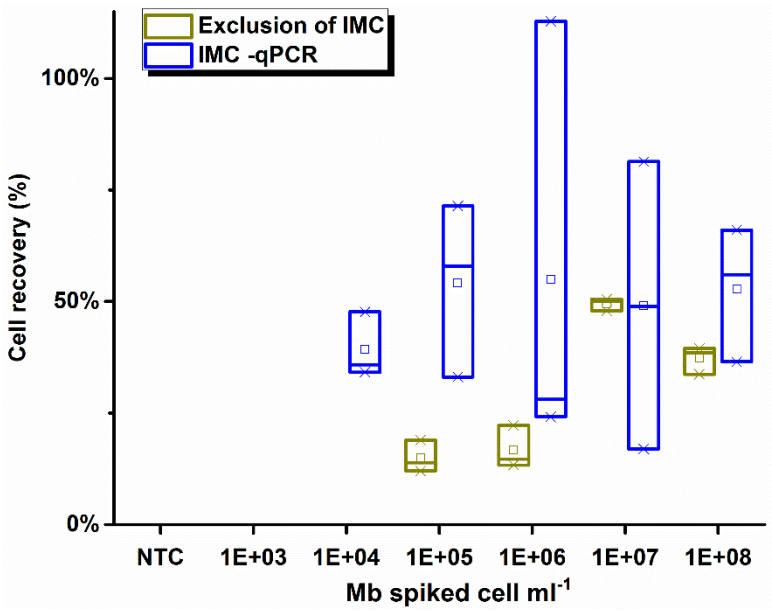


Figure 3.11 Comparison of efficiency for Mb capture using IMC-qPCR and qPCR methods with spiked cells in raw milk samples.

3.3.4. qPCR and internal control: avoidance of false negative

Inhibition of PCR reaction can result in false negative this can be due to arrange of inhibitor concentrated during DNA extraction.

An internal control based on an exotic gene fragment from green fluorescent protein flanked by RD4 scar priming sites was used previously but not in the same reaction mixture (Pontiroli, Travis et al. 2011). Therefore, experiments for detection of Mb were done to improve the RD4 scar method by incorporating internal control plasmid in with the sample and multiplexing the two, saving time and money. These experiments aimed to determine if the internal control as target for the same primers competed with the sample target and reduced sensitivity. The high specificity of plasmid using RD4 scar qPCR was reported in the same study (Pontiroli, Travis et al. 2011).

The comparison was performed between a set of log10 serial dilutions of Mb as reference with and without the internal control plasmid. The Figure 3.12 shows the obvious curve shifted from 30 to 50 C_T with threshold detection while the high concentration of plasmid was exposed to Mb reference dilution. The threshold value used in this study was 0.02 ΔRn (logarithm of normalised reporter) which was plotted against cycle time to establish an Mb reference amplification plot. The genome equivalents of the samples was evaluated based on amplification plot in qPCR. Therefore, the high concentration samples were amplified prior to low titre ones. It was ascertained that the plasmid added in the same PCR reaction tube with Mb reference triggered competition because of shared amplification target (Figure 3.12).

The noticeable curve shift in C_T was revealed when high titre plasmid was added and amplified with Mb reference together (Figure 3.12).

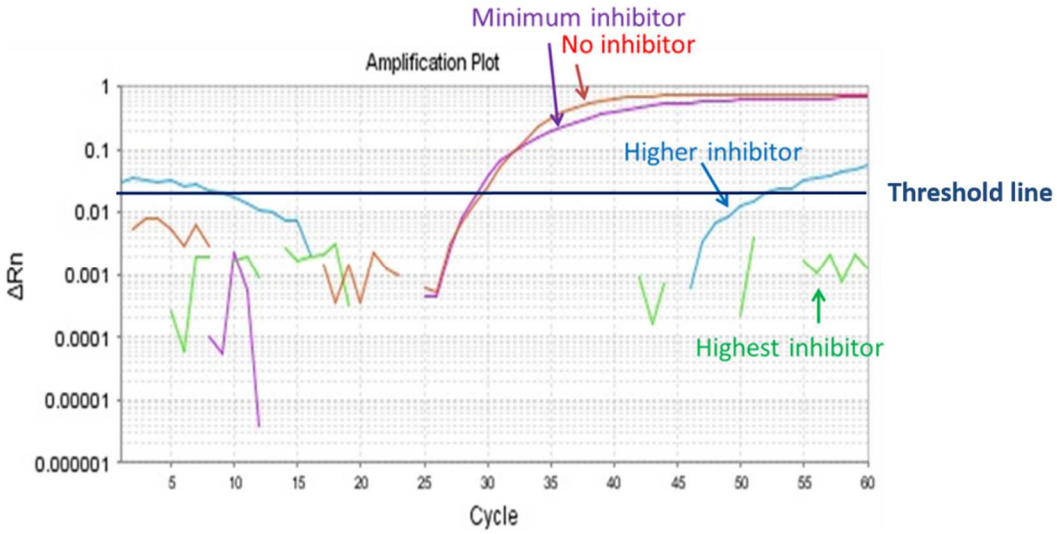


Figure 3.12 Effects of added internal control plasmid on specific Mb detection using the RD4 scar qPCR. The ΔRn of CT for in inhibition is summarised in Table 3.1.

Table 3.1 Summary of the ΔRn of CT for in inhibition.

Inhibition	ΔRn of CT
No inhibition	29.71
Minimum inhibition	29.47
Higher inhibition	52.35
Highest inhibition	0

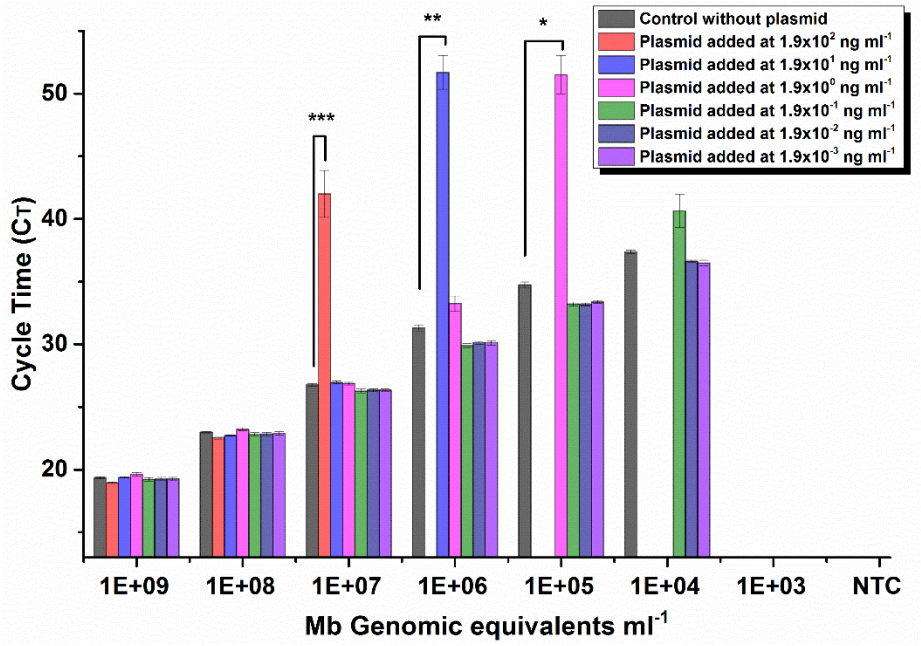


Figure 3.13 Effect of varying levels of inhibition control plasmid on cycle time (C_T) for detection of Mb DNA using RD4 scar qPCR assay

The varying log10 serial dilution of plasmid ranged from $1.9x10^{2}$ to $1.9x10^{-3}$ ng ml^{-1} and was incorporated in the whole set of Mb reference samples which were tested using qPCR with RD4 scar assay (Figure 3.13). Failure to detect Mb occurred in a number of assays for example at 10^{6} Mb with $1.9x10^{2}$ plasmid concentration. Therefore it was reported that the internal control was included in the test at a concentration of $1.9x10^{-1}$ ng ml^{-1}. The limit of detection for Mb was 10^{4} genome copies ml^{-1} with the level of plasmid template at $1.9x10^{-1}$ ng ml^{-1} achieved highest sensitivity.

3.4. Discussion

The DNA isolation kits were selected from different commercial and manual methods used on environmental samples including liquid and solid samples. The FastDNA isolation kit was more effective than DNeasy Tissue & blood kit on isolation of DNA from liquid culture and achieved $3.90x10^9$ genome copies ml^{-1} on average of triplicates. It is assumed that the combined lysis method consisted of mechanical and chemical lysis approaches which were more efficient than just cells disruption for extraction of DNA.

A previous study compared different methods for DNA extraction from raw milk and recommend the PowerFood kit (Quigley, O'Sullivan et al. 2012). DNA yield was 909.53 ± 6.0 ng ml^{-1} with 1.85 A260/280 nm purity using PowerFood kit compared to the second best commercial kit which was "Milk Bacterial Isolation kit" which yielded 835.96 ± 57.29 ng ml^{-1} DNA with 1.56 purity and the best manual method among the three tested was the Lytic method which extracted 776.42 ± 25.5 ng ml^{-1} with 1.23 purity. In addition the PCR yield was also compared after DNA extraction using kits and the PowerFood kit achieved 5132.86 ± 77.47 ng per PCR run which was slightly lower than the optimal Lytic method which achieved 5143.43 ± 62.97 ng per PCR run (Quigley, O'Sullivan et al. 2012). The same result was achieved in the current study using O'Neil Method and blood & tissue kit.

The specificity and sensitivity of each set of primers and probes provided in this study was tested for detection of Mtb and Mb cells from liquid and solid samples. The Wbbl1 primer was reported to identify Mtb and *M. canettii* as these species share the same RD9 target (Reddington, O'Grady et al. 2011). However the Wbbl1 primers

targets the region of the *wbbl1* gene in Mtb, *M. canettii* and *M. africanum* which encodes rhamnosyl transferase, and enzyme that insert rhamnose into the cell wall and this step is essential for viability (Ma, Pan et al. 2002). In contrast to Wbbl1, the RD9 primer only detected Mtb and *M. canettii* (Halse, Escuyer et al. 2011). The RD9 primer therefore was chosen for Mtb detection because it had higher specificity and sensitivity than the Wbbl1 primer in the trial from liquid samples. This is the first report using RD9 primer in both liquid and solid environmental samples.

The LepA primer was efficient and specific for detection of MTBC members from culture, no other sample type was tested (Reddington, O'Grady et al. 2011) . The RD4 scar primer was compared to LepA and found to be more specific for detection of Mb only (Reddington, O'Grady et al. 2011) and more sensitive than LepA. The RD4 scar primer therefore was selected as qPCR primer for Mb in further environmental sample screening.

However the sensitivity in raw milk was affected by PCR inhibitors. Therefore the efficiency for limit of detection on Mtb and Mb cells was 10^5 genome copies ml^{-1}. The development of an antibody based approach was valuable and successfully addressed the issue of PCR inhibitors contained in raw milk by concentration and purification of target cells prior to DNA extraction.

IMC methods have been applied with large number of bacterial species for specific cell capture (Corner, John et al. 1988, Wipat, Wellington et al. 1994, Sweeney, Courtenay et al. 2006, Grant and Stewart 2015). MAbs have been for selective recovery of Mtb from sputum smears plated on agar (Corner, John et al. 1988). Methods had been developed to rescue species from soil (Wipat, Wellington et al.

1994). In addition MAbs yielded high specificity and sensitivity for Mb detection in smear, faecal and tissue samples (Sweeney, Courtenay et al. 2006, Grant and Stewart 2015).

Dynabeads had been used for both direct immunomagnetic separation (IMS) and with a phage coating (Stewart, McNair et al. 2012). The specificity and sensitivity of IMS was similar to the conventional IMC but IMS reduced the time and budget for route diagnosis and were possible to use on tissue samples (Stewart, McNair et al. 2012). In the current study, over 70 % capture was achieved without phage attachment coating as the Dynabeads performed well with antibody.

The high specificity of MAbs against Mb was reported for detection of Mb cells in environmental samples (Corner, John et al. 1988, Wipat, Wellington et al. 1994, Sweeney, Courtenay et al. 2006, Grant and Stewart 2015) but MAbs demonstrated lower cell recovery rate compared to PAbs raised against Mtb in this study. The possibility of reduced sensitivity for the MAbs was possibly due to being held at -20 °C for a year. The PAbs applied in IMC were therefore selected compared to MAbs for liquid environmental sample screening.

The sensitivity was enhanced using IMC for recovery from milk, however, this antibody based approach is reliable but laborious, expensive and unsuitable for high throughput and the majority of samples.

Several compounds can act as PCR inhibitors such as humic acid in soil, biliary salts, urea, haemoglobin and heparin in faeces and metal ions in milk (Section 3.5) (Lantz PG 1997, Wilson 1997, Jost 2007). The inhibition control assay was first reported as

an efficient method for identification of false negatives using a separate assay to avoid primer competition (Pontiroli, Travis et al. 2011). However the current study demonstrated that this can be avoided by the low primer concentration in the same reaction.

In conclusion, the FastDNA isolation kit, PowerWater and PowerFood isolation kits are recommended for DNA extraction in solids, water filters and milk samples respectively to ensure reliability and sensitivity to quantify the pathogen. The RD9 and RD4 scar primers and probes are considered as the major tools for identification of Mtb and Mb due to their sensitivity and specificity within diverse environmental samples. In addition IMC helps to improve sensitivity in liquid samples but further validation is required for solid samples. The inhibition control assay enabled reliable and sensitive detection of false negatives and was applied for environmental sample screening in the current study.

Chapter 4 Prevalence of the pathogens Mtb and Mb in environmental samples with relevance to epidemiology of tuberculosis.

4.1. Introduction

The region proposed in the current study was the Iringa administrative region near the Ruaha, Tanzania's largest national park. This park is one of Africa's wildlife protected areas but it has suffered from water scarcity of the Great Ruaha River since 1993 due to uncontrolled agricultural water diversions and intensive livestock grazing in wetlands. The result of this water scarcity means a rising wildlife-agriculture conflict in areas of human settlement adjacent to the Ruaha protected area. There is then the possibility of disease spread within livestock via the same environmental reservoirs shared between wildlife and livestock animals. The study site was selected to cover the Iringa area of overlap between the village and the park due to this influence in the Ruaha (Figure 4.1).

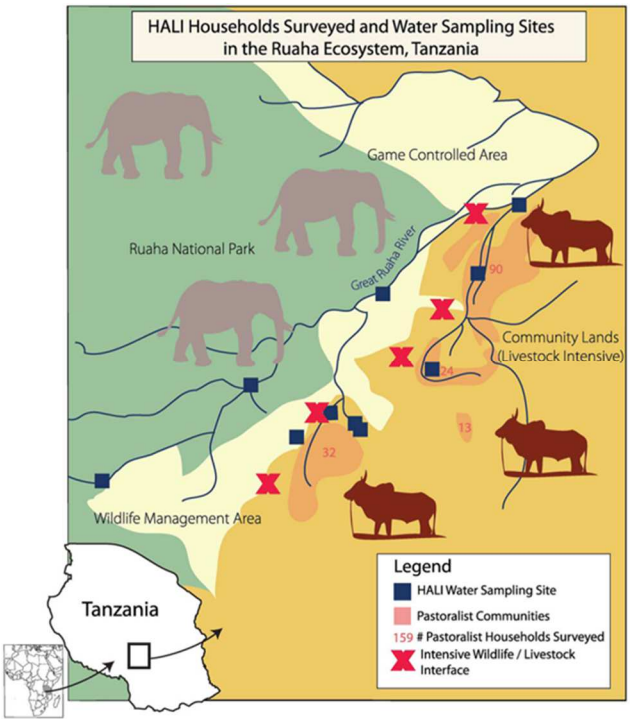


Figure 4.1 Ruaha study region where preliminary studies were conducted in Tanzania.

The Iringa region is in South-Central Tanzania, which has a relatively fragmented human population that is clustered in small villages, with livestock-raising and farming as core livelihoods of pastoralists and agriculturalists respectively. The pastoralists often lived in very close contact with their livestock animals and consume milk as well as animal products. Many pastoralists migrated extensively with their livestock in order to access food and water sources, while the agriculturalists tended to cultivate in the same area without traveling. Both HIV and TB were endemic in the area where the pastoralist and agriculturalists live. A previous study hypothesised that the pastoralists would have a higher prevalence of TB compared to agriculturalists due to living in close contact with their animals and consumption of animal products; however, there is no substantial evidence supporting this hypothesis (Hopewell 2011). Another study also indicated a further mycobacterial pathogen, MAP, which was reported to have contaminated the surrounding cattle barn via bio-aerosols from cattle shedding MAP (Eisenberg, Nielen et al. 2010). Previous and current studies have depicted the high density of Mb cells present in soil at wildlife animal badger setts (Young, Gormley et al. 2005, Hayley King 2015). A recent study compared the dominant Mtb genotype in TB patients with that of Mtb isolated from soil and water sources indicating shedding and environmental survival (Velayati, Farnia et al. 2015). These studies all assumed the environment was capable of being a potential role in SGM transmission via soil, water and animal shedding.

The experiment was designed in order to prove the hypothesis that Mtb and Mb are shed into the environment and survive; that pastoralists have a greater exposure to Mb than agriculturalists. The first task was to collect environmental samples from

pastoralist villages in the study region for detection of Mb and Mtb species using molecular approaches and R-mapping to achieve a spatial study. The correlation was then built between prevalence of Mtb and Mb in relation to location and the metadata collected.

A previous study indicated that identification and differentiation of each MTBC individual provide a significant challenge because of high similarity between Mtb and Mb can approaches 99.95% in whole genome nucleotide level and identical highly conserved 16S rRNA region (Garnier, Eiglmeier et al. 2003, Tortoli 2012).

In Chapter 3 it was shown that it was possible to use molecular techniques for the detection of Mtb and Mb respectively from environmental samples. These approaches consisted of efficient DNA isolation kits, the specific primers and probes for Mtb and Mb based on regions of deletions. For quantification qPCR, IMC for sensitivity enhanced capture and internal control assay to avoid false negatives all contributed to highly sensitive and specific assays. The methods were applied for detection of both Mb and Mtb in environmental samples collected in Tanzania as part of a wider study on the epidemiology, risk factors and transmission dynamics of Mtb and Mb between humans and animals in a well-defined rural population in Tanzania. The quantification of these two species was calculated and statistical analyses were conducted to compare the genome equivalents presented in the liquid samples collected during the wet and dry seasons respectively. The study region was in the Iringa area close to the Ruaha national park; six pastoralist villages were selected as sample sites, each village having five or more households and collectively per village approximately 100 or more cattle. The majority also kept goats herds of 20-50

animals with their kids. The location of each village shows the spread of sample sites over a region of interface between livestock rearing and wildlife (Figure 4.2). The distribution of water and sediment sample collection was based on watering holes used by the selected villages (Figure 4.3). More information about each village is given in Table 2.2 and 2.3.

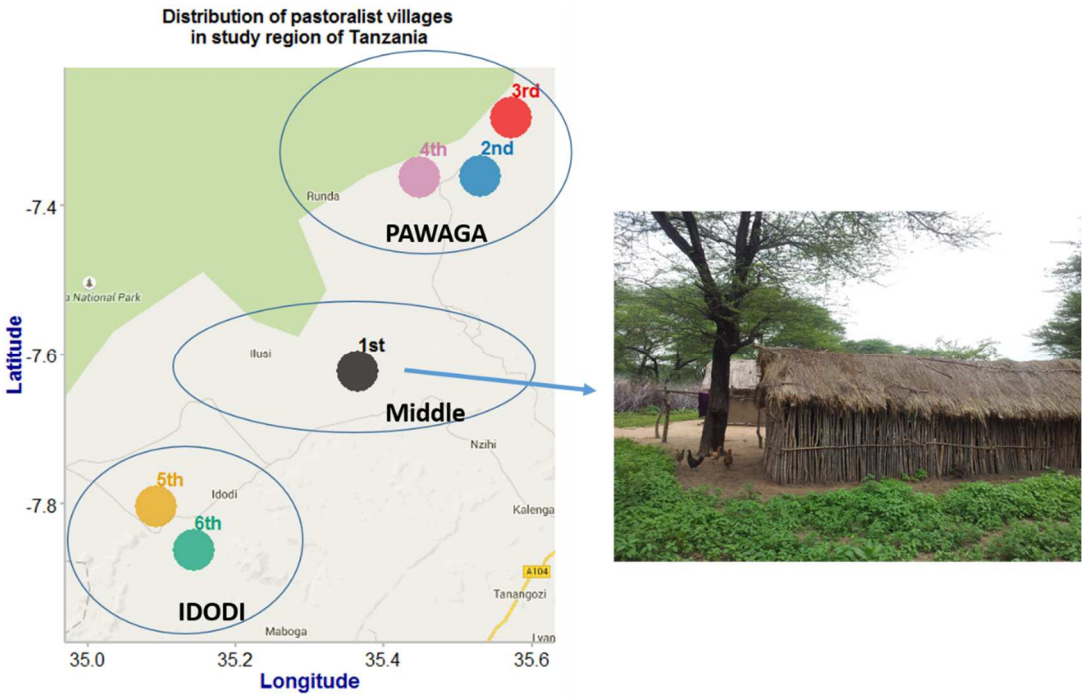


Figure 4.2 Distribution of six pastoralist villages in the study region of Tanzania. The green area indicates the location of Ruaha national park.

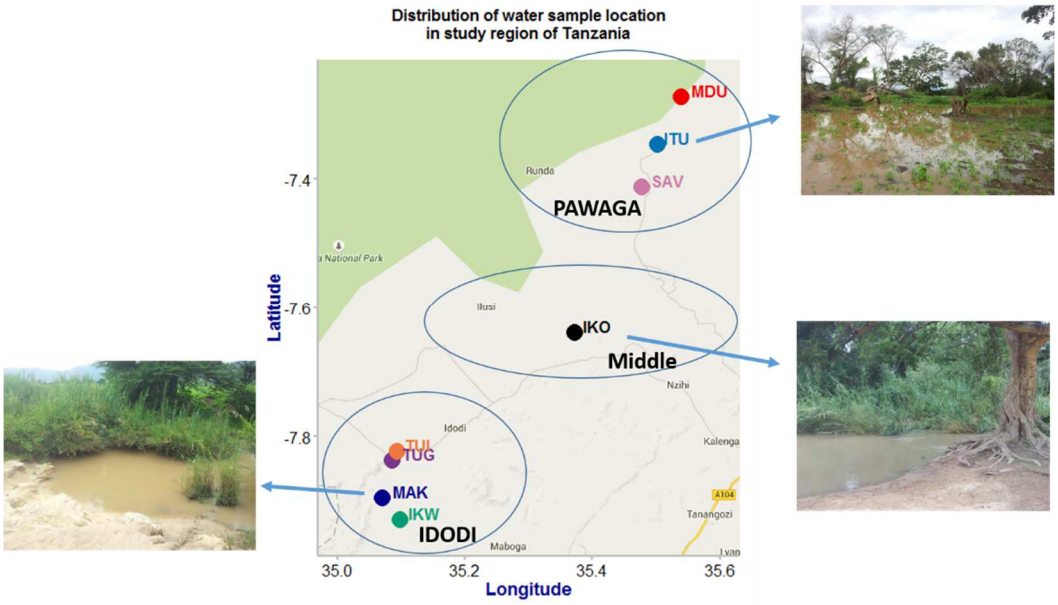


Figure 4.3 Distribution of water and sediment sample location in the study region of Tanzania. The green area indicates the location of Ruaha national park.

There were 803 samples in total collected from the dry season in 2012 compared to 719 in the wet season in 2014 (Table 4.1). The samples consisted of faeces, boma soil, household dust, water and sediment. All solid samples including faeces, soil, dust and sediment were extracted using FastDNA SPIN kit and water filter samples were processed using PowerWater kit for DNA isolation. The qPCR was applied with specific primers and probes, RD4 scar and RD9 for detection of Mb and Mtb respectively and statistical analysis was performed by OriginLab.

Table 4.1 Number of diverse environmental samples collected in the dry season of 2012 and the wet season of 2014

Samples	Dry season	Wet season
Cowpat and goat faeces	394 116	370 96
Boma soil	171	142
Household dust	90	75
Water	16 / 6	18 / 9
Sediment	16 / 6	18 / 9

4.2. Aims

1. To quantify the number of Mtb and Mb contained in Tanzanian environmental samples and compare prevalence in the wet and dry season samples.

2. To introduce spatial analysis to investigate the distribution of TB and bTB in the region.

3. To estimate the correlation between environmental prevalence and metadata including human activities, herd size, season and other variable (Table 2.2 and 2.3).

4.3. Results

4.3.1. Quantification of Mtb and Mb in environmental samples

The RD4 scar and RD9 primers provided sufficient specificity and sensitivity (Chapter 3) for detection of Mb and Mtb respectively. All environmental samples were tested using qPCR with these specific RD probes to quantify the genome number according to a tenfold dilution series from 10^6 to 1 genome equivalents μl^{-1} .The prevalence of

target genome equivalents was compared for samples taken in the wet and dry seasons.

4.3.1.1. Quantification of Mtb and Mb in faecal samples

The 976 faecal samples consisting of 764 cattle and 212 goat samples were examined overall with approximately 60 cattle and 20 goat samples taken from each village. The Mb prevalence in the cattle faeces from different villages indicated high prevalence in both wet and dry season in village 6 which was significantly higher in the wet season but overall counts were higher in the dry season across the majority of villages (Figure 4.4). A similar trend was seen for Mb prevalence in goat faeces with village 6 again showing highest counts in both wet and dry season. The box and whisker plots showed the ranges of Mb genomic equivalents detected from each sample and the outliers represented the most extreme observations including sample maximum or minimum.

The negative skewed distribution of Mb quantification in cattle faecal samples in the dry season was significantly higher (P-value < 0.001) than narrow distribution in the wet season which showed a positive skew (Figure 4.5). However there was no significant difference (P-value > 0.05) in terms of Mb genomic equivalents in goat faeces between two seasons (Figure 4.6).

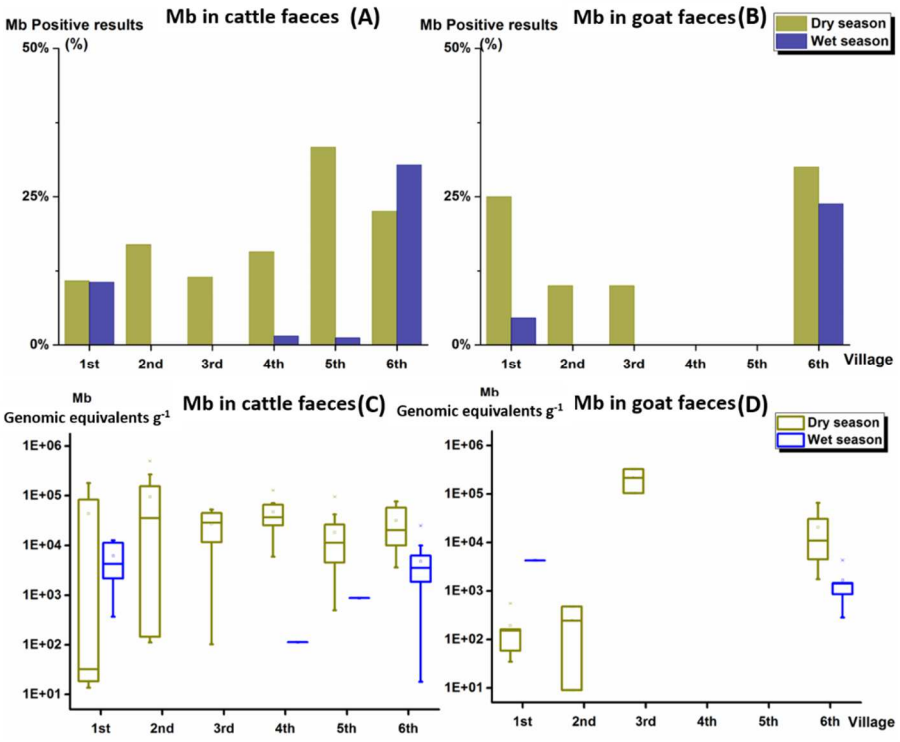


Figure 4.4 Comparison of Mb positives (A, B) and quantification of Mb (C, D) in each village over two seasons in the cattle (A, C) and goat faeces (B, D).

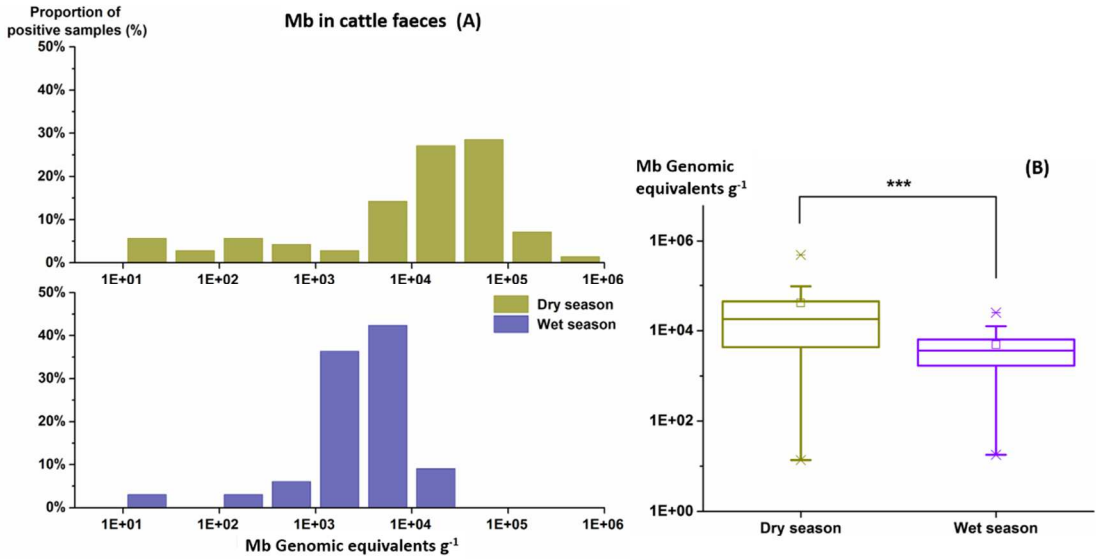


Figure 4.5 Mb prevalence in the cattle faeces (A) and ranges (B) over two seasons. ***Mann Whitney test (P-value < 0.001).

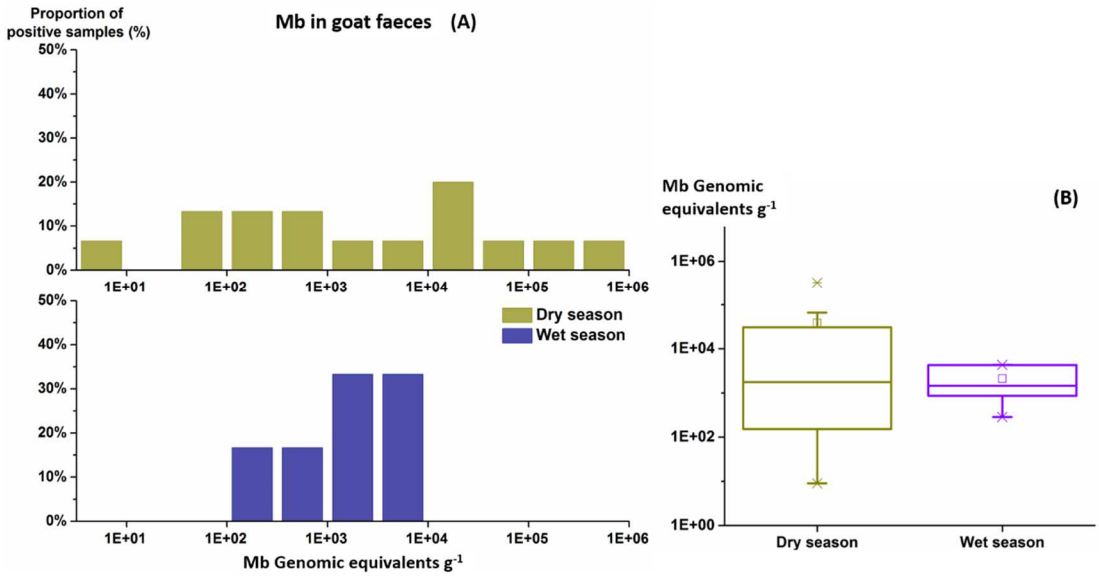


Figure 4.6 Mb prevalence in the goat faeces (A) and ranges (B) over two seasons.

The Mtb from cattle faecal samples in village 1 was highest proportion of any dataset for both cattle and goats across all seasons and all villages with over 50 % positive in the dry season (Figure 4.7). However Mtb in faecal samples in village 5 were significantly high in the wet season because Mtb was rarely detected in faeces (Figure 4.7). There was a pronounced negative skew on the wet season counts for cattle faeces as the dry season counts although with a similar negative skew were approximately 50 % lower (P-value < 0.05) (Figure 4.8).

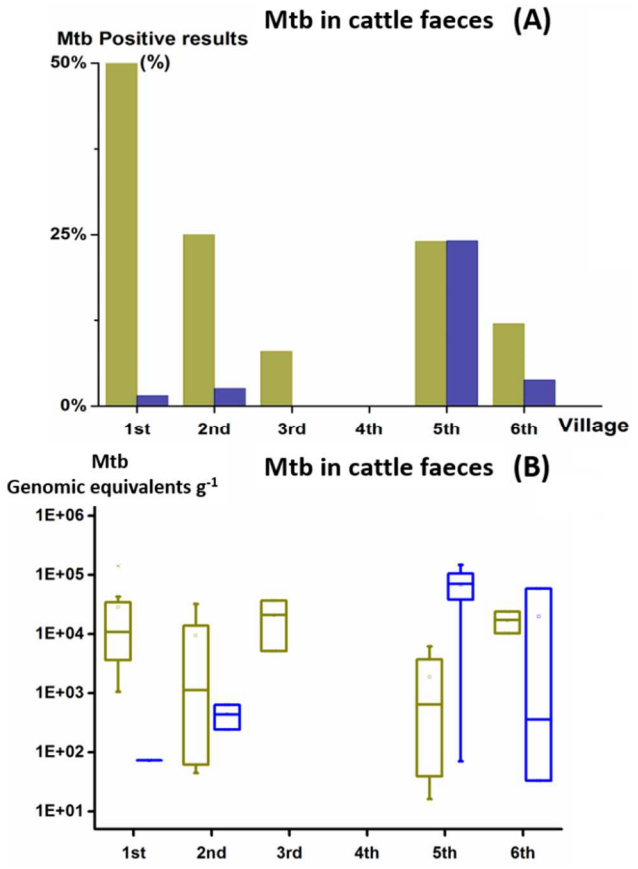


Figure 4.7 Comparison of Mtb positive results (A) and quantification of Mtb (B) in each village in the cattle over two seasons. No detection of Mtb in goat faeces.

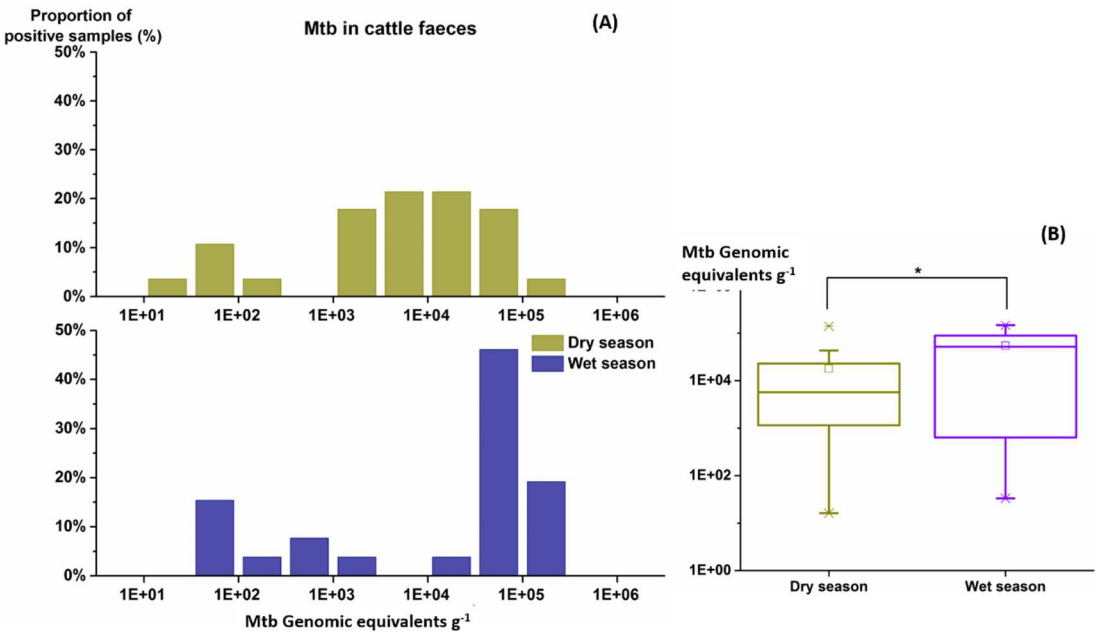


Figure 4.8 Mtb prevalence in the cattle faeces (A) and ranges (B) over two seasons. *Mann Whitney test (P-value < 0.05).

4.3.1.2. Quantification of Mb and Mtb in boma soil samples

There were very few Mb positives in both cattle and goat boma soil and Mb positives only present in village 1 and 6 and only in the wet season (Figure 4.9). No Mb positives from the dry season were detected in any pastoralist villages. The Mb positives ranged between 10^2 to 10^4 genomic equivalents g^{-1} in both cattle and goat soil samples (Figure 4.10 and 4.11).

Mtb was only detected in village 1 in cattle boma soil in the wet season and none was found in the dry season. As for the faeces no Mtb was detected in the goat boma soil (Figure 4.12).

The moisture contained in the boma soil was measured when samples were taken and was significantly higher (P-value < 0.001) in the wet season in both cattle and goat soil samples (Figure 4.13).

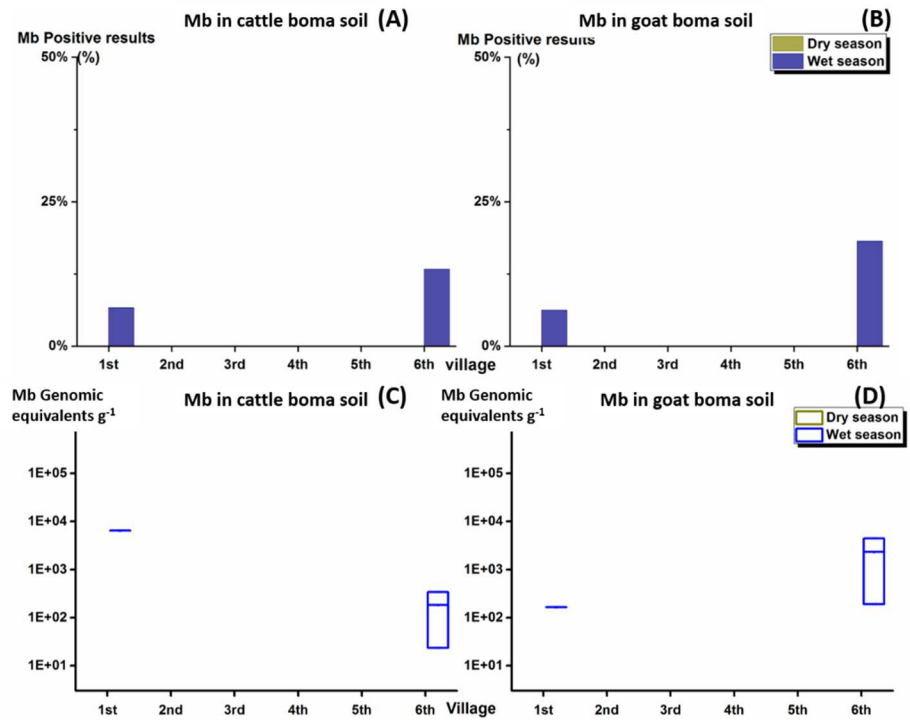


Figure 4.9 Comparison of Mb positives (A, B) and quantification of Mb (C, D) in each village over two seasons for cattle (A, C) and goat boma soil (B, D).

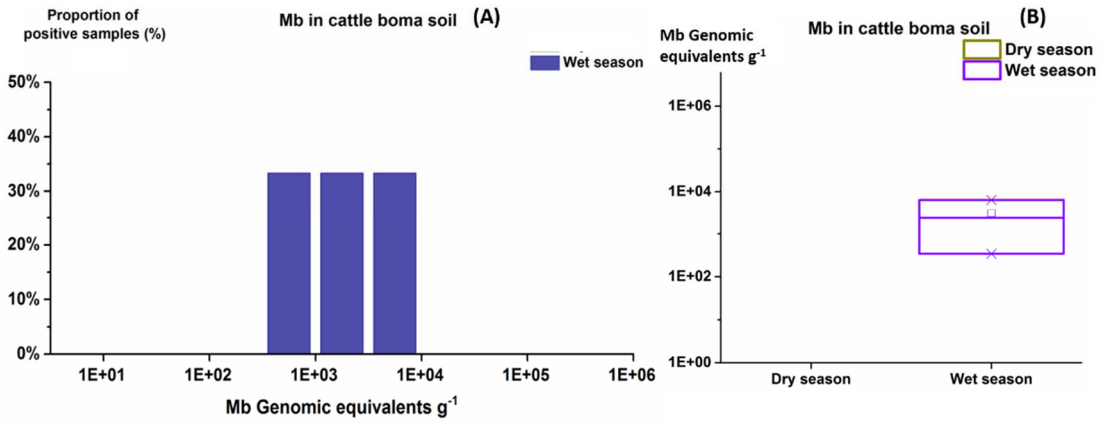


Figure 4.10 Mb prevalence (A) and the range (B) in cattle boma soil over two seasons. No Mb detection in the dry season.

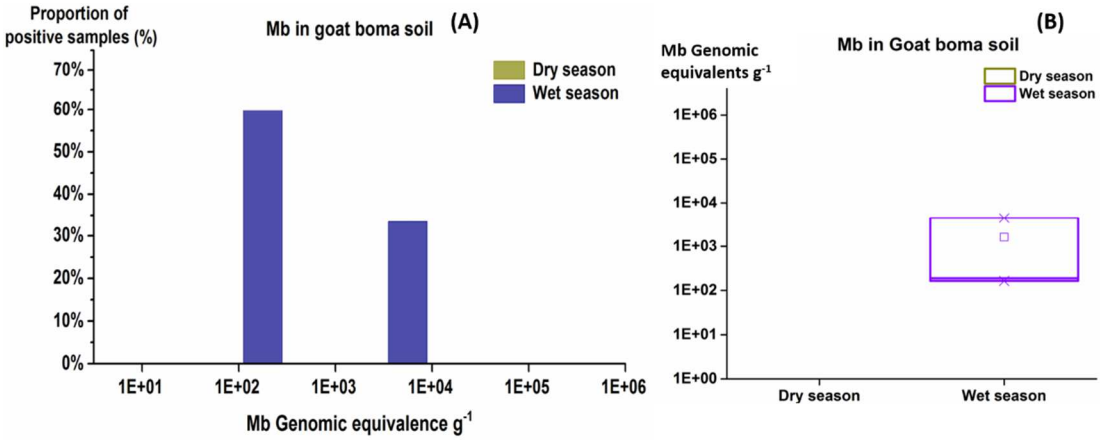


Figure 4.11 Mb prevalence (A) and the range (B) in the goat faeces over two seasons. No Mb detection in the dry season.

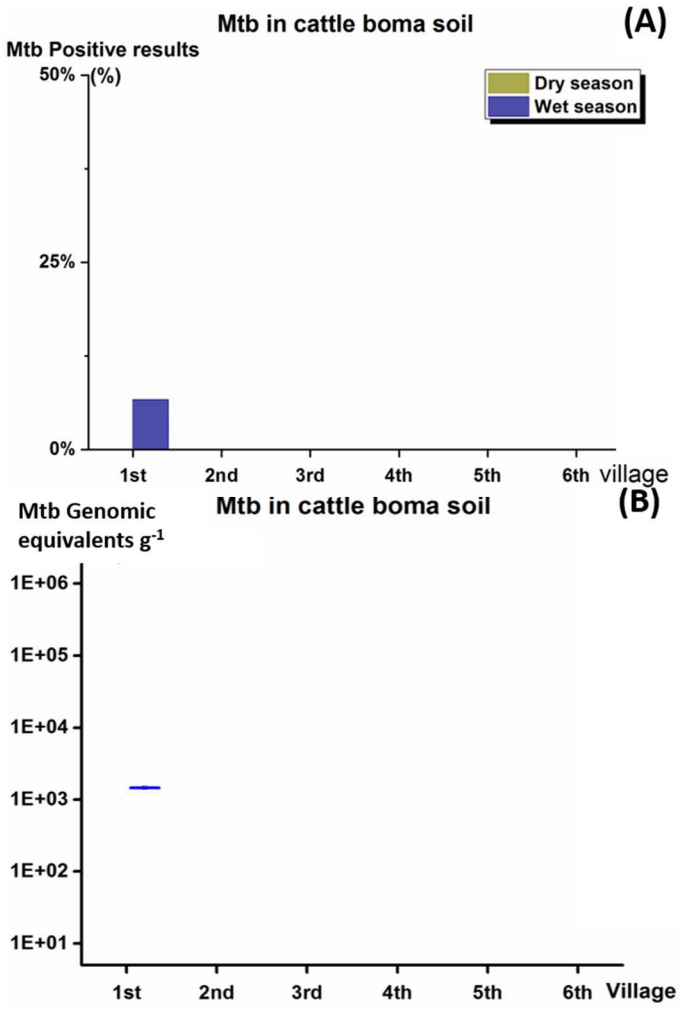


Figure 4.12 Mtb prevalence (A) and quantification (B) in each village over two seasons in cattle boma soil. No Mtb detection in goat boma soil samples.

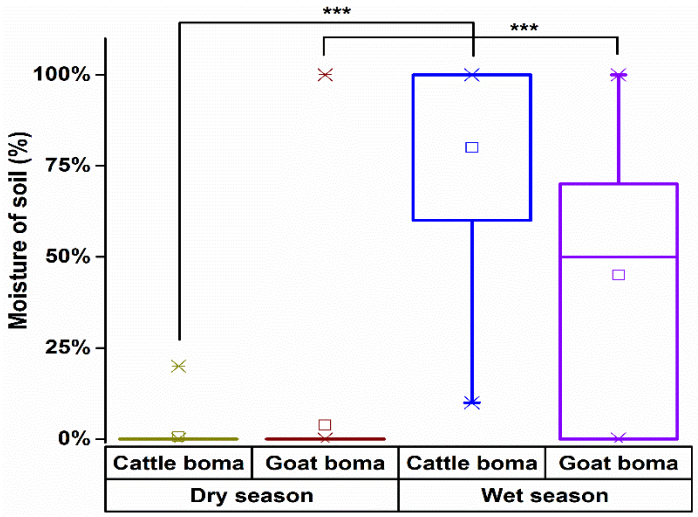


Figure 4.13 Moisture content in the cattle and goat boma soil over two seasons ***Mann Whitney (P-value < 0.001).

4.3.1.3. Quantification of Mb and Mtb in household dust samples

Mb was rarely detected from household dust and detected only village 4 in the dry season (Figure 4.14). Village 5 showed significant prevalence of Mtb in both seasons (Figure 4.15) where positives detected in the wet season reached approximately 10^6 genomic equivalents g^{-1} (Figure 4.16).

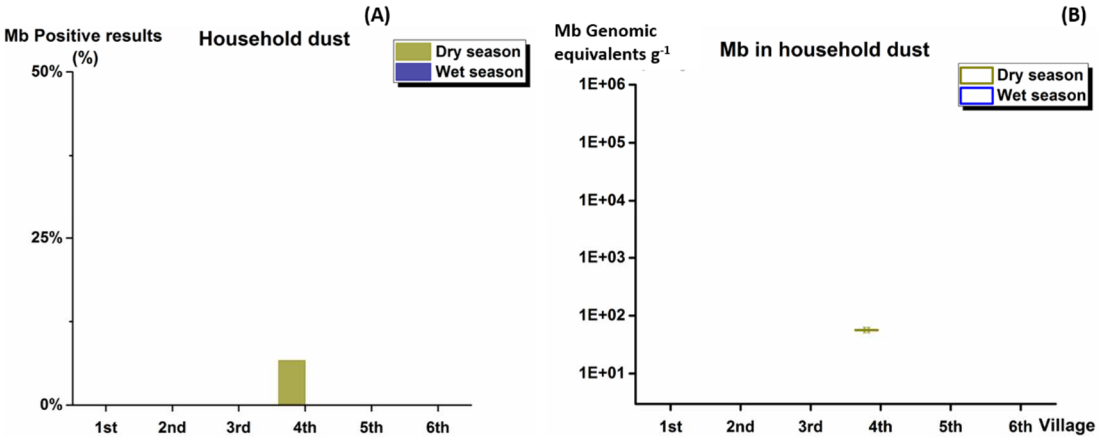


Figure 4.14 Comparison of Mb positive results (A) and quantification (B) in each village over two seasons in the household dust.

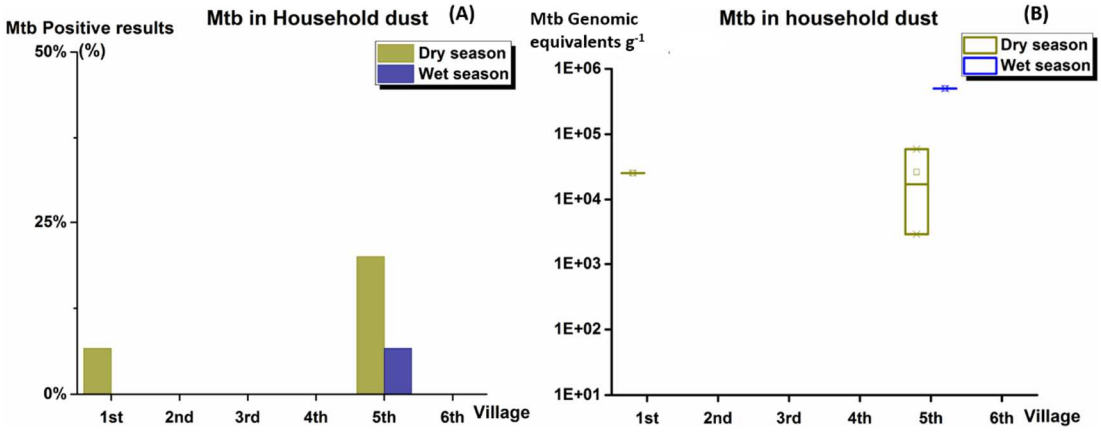


Figure 4.15 Mtb positive results (A) and quantification (B) in each village over two seasons in the household dust samples.

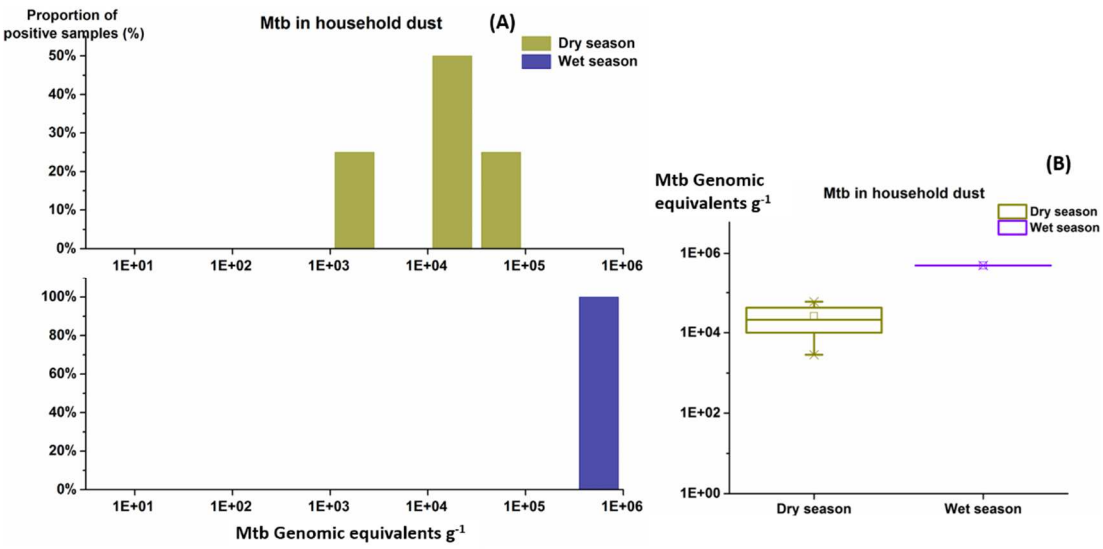


Figure 4.16 Mtb prevalence in the household dust (A) and the range (B) over two seasons

4.3.1.4. Quantification of Mtb and Mb in water and sediment samples

A high prevalence of Mb was observed in water filters in the wet season (Figure 4.17) as nearly 80 % of water samples contained Mb between 10^3 to 10^4 gene copies filter^{-1} (Figure 4.18). Nevertheless, Mb species was below qPCR detection in all sediment samples.

The Mtb species was identified in only one sediment sample from eight river locations, Ikwavila (IKW) and Ikonongo (IKO) at approximately 10^4 Mtb genome equivalents filter^{-1} (Figure 4.19). However in the filtered water samples Mtb pathogens were under the detection limit for qPCR from water samples in both seasons.

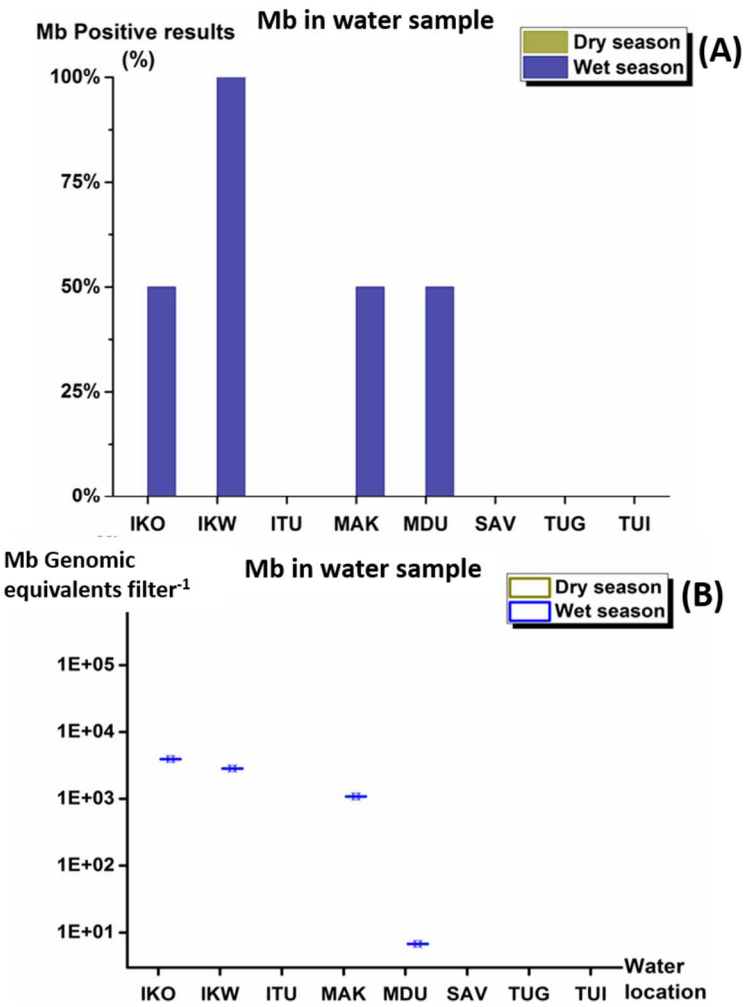


Figure 4.17 Comparison of Mb positive results (A) and quantification (B) in water filters. No Mb detection in sediment samples.

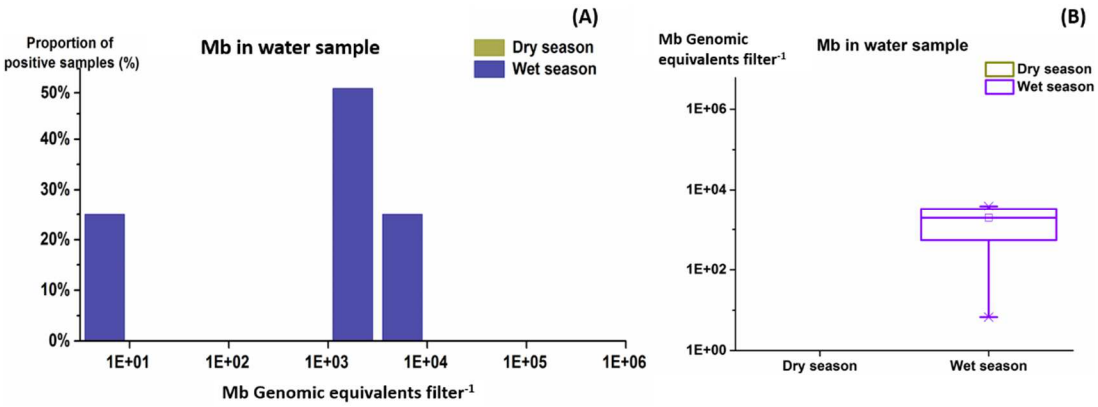


Figure 4.18 Mb prevalence in the water filters (A) and the range (B) within the two seasons. No Mb detection in the dry season.

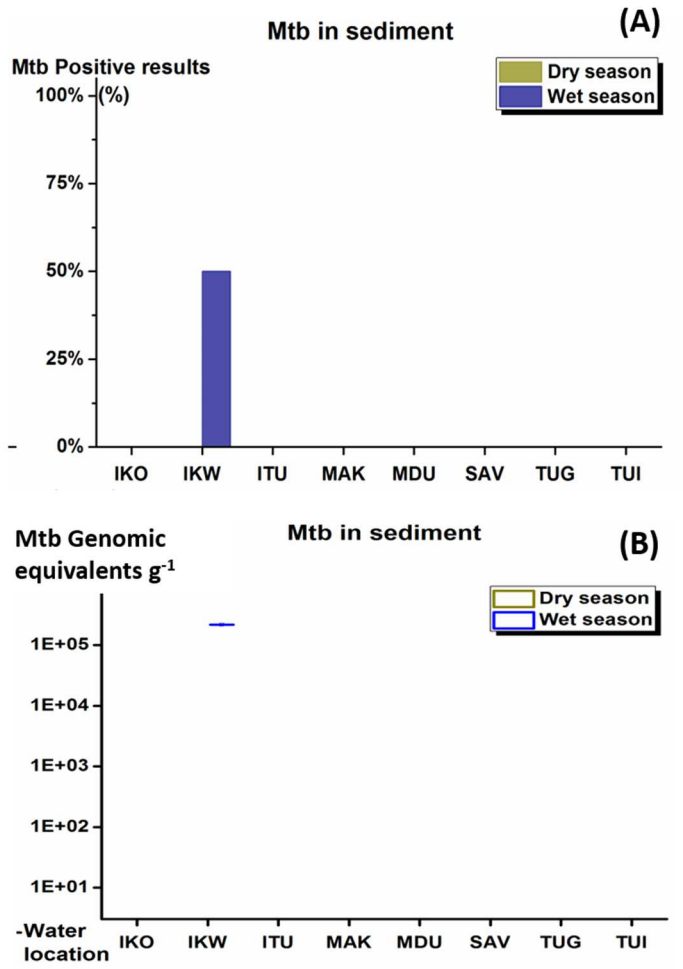


Figure 4.19 Comparison of Mtb positive results (A) and quantification (B) in each river sample location over two seasons in sediment samples. No Mtb detection in water filters.

4.3.2. Spatial analysis correlations with Mtb and Mb abundance

The prevalence of Mb and Mtb was revealed using the R programme combined with google map to describe the distribution of Mtb and Mb shedding positives in each household from the six villages.

4.3.2.1. Prevalence of Mb in the environmental samples

The bubble plot (Figure 4.20) demonstrates the spatial relationship between sample sites and Mb shedding into the environment via cattle faeces in the wet and dry seasons. The high prevalence in villages 5 and 6 can be attributed to herds from several homesteads but more than 70 % of Mb prevalence in cattle faeces was attributed to one homestead in each village in the dry season. In the wet season the majority of shedding was from one or two homesteads particularly again in village 6.

The high Mb prevalence via shedding as goat faeces was primarily attributed to villages 1 and 6 in both seasons (Figure 4.21). High levels of shedding across both seasons in village 6 indicated considerable infection within the livestock of this area but primarily focused on two homesteads.

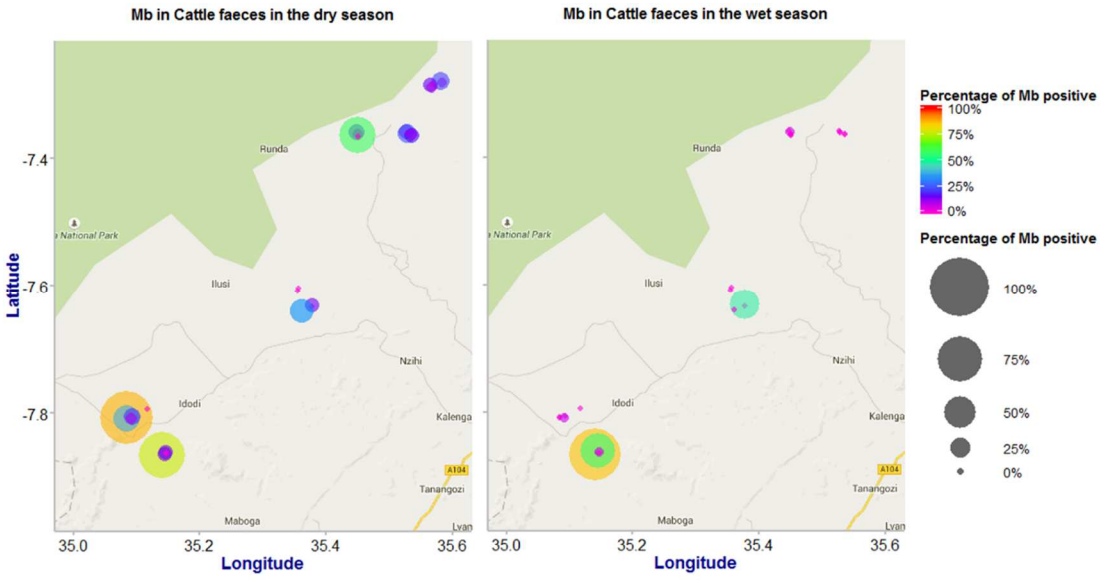


Figure 4.20 Mb incidence in cattle faeces in each homestead over two seasons. Scale indicated as heat map and bubble pot to avoid overlaps.

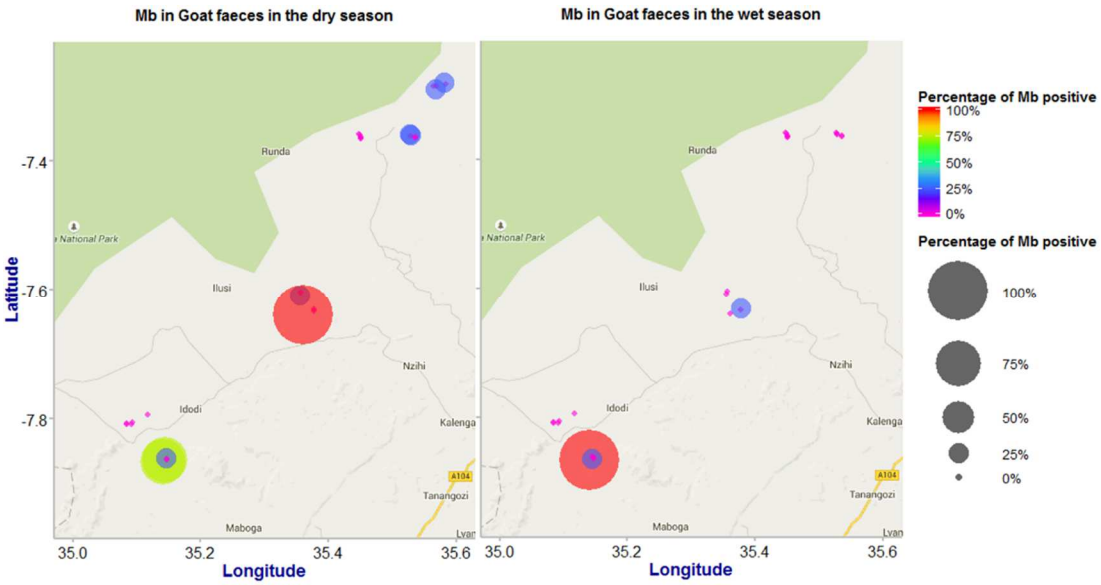


Figure 4.21 Mb incidence in goat faeces in each homestead over two seasons. Scale indicated as heat map and bubble pot to avoid overlaps.

There were very few Mb positives in boma soil with only one homestead in village 1 showed Mb positives in contrast to two homesteads in village 6 in the wet season but none in the dry season (Figure 4.22). The majority of Mb positives were similar between cattle and goat soil from village 1 and 6 but one homestead in village 6 (as before) proved a hotspot for shedding in goat boma soil (Figure 4.23).

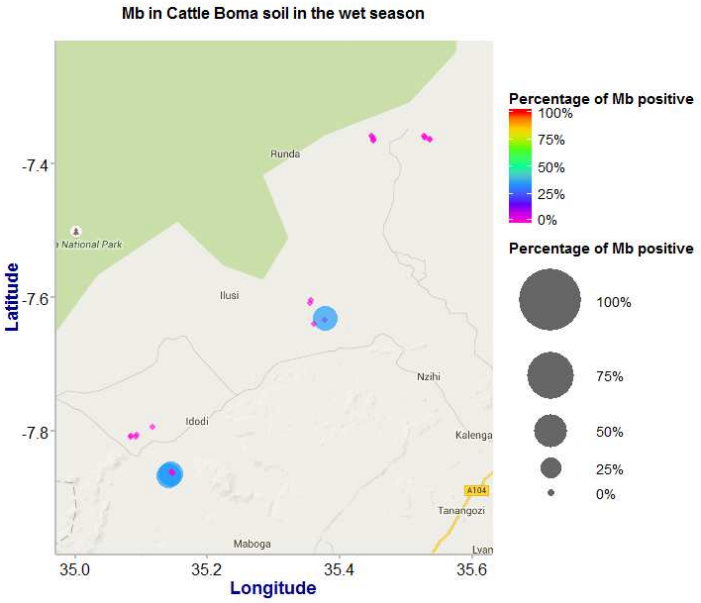


Figure 4.22 Mb incidence in cattle boma soil in each homestead in the wet season. No Mb detection in the dry season.

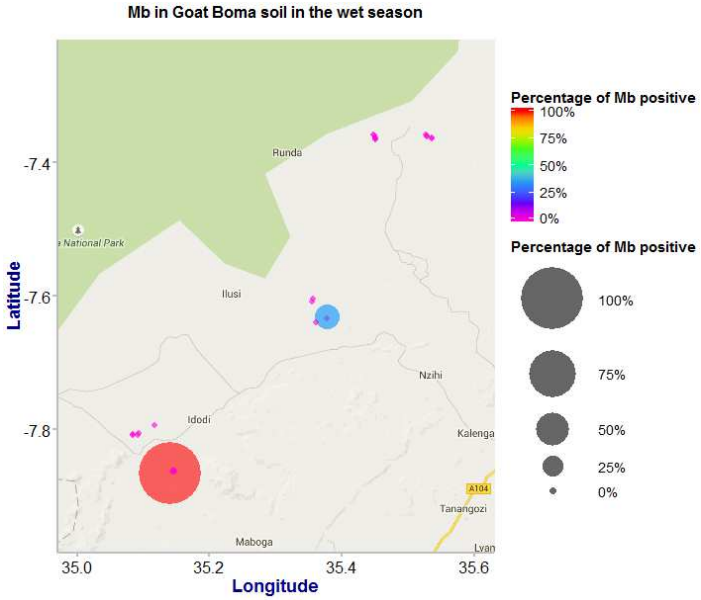


Figure 4.23 Mb incidence in goat boma soil in each homestead in the wet season. No Mb detection in the dry season.

Despite extensive Mb shedding via cattle and goat faeces there was negligible contamination of household dust with Mb in the dry or wet season. (Figure 4.24).

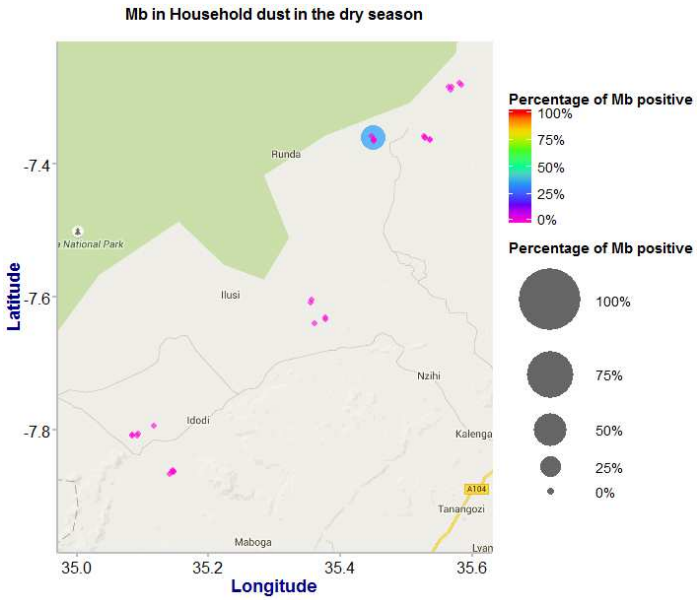


Figure 4.24 Mb incidence in dust within individual households in the dry season. No Mb detection in the wet season.

The Figure 4.25 and appendix A.1 represents the Mb prevalence in water and sediment samples and Mb was under detection of limit in the dry season. It is worth noting that the positive water filters are from river site close to village 1 and 6 with heavy Mb shedding data.

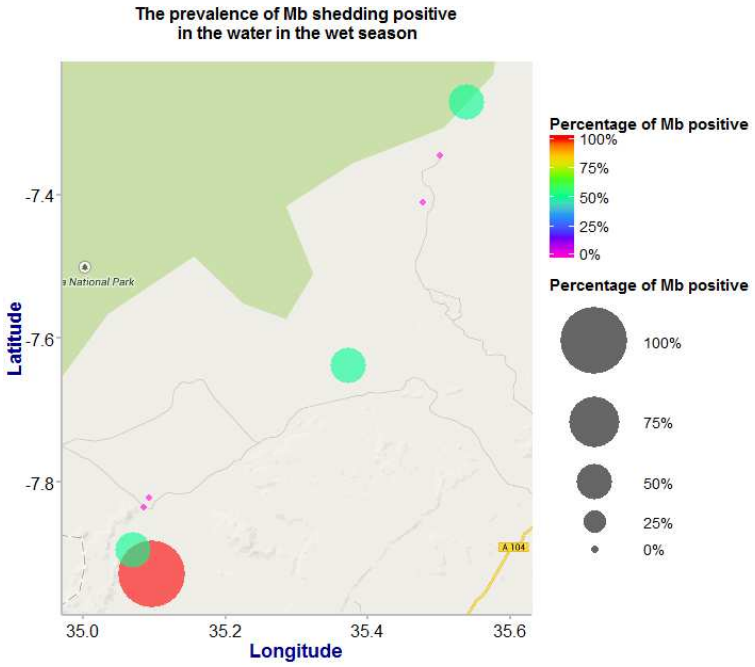


Figure 4.25 Mb incidence in water taken from eight river locations along the River Ruaha. No Mb detection in the dry season.

4.3.2.2. Prevalence of Mtb in the environmental samples

Mtb was shed into the environment via cattle faeces in the wet and dry seasons. The high prevalence in villages 5 can be attributed to herds from several homesteads but more than 70 % of Mtb prevalence in cattle faeces was attributed to one homestead in village 5 in each wet and dry season. In the wet season the majority of shedding was from two or three homesteads particularly in villages 1, 2, 5 and 6 (Figure 4.26). Mtb positives was under limit of detection via goat faeces in both seasons (Figure A1.3).

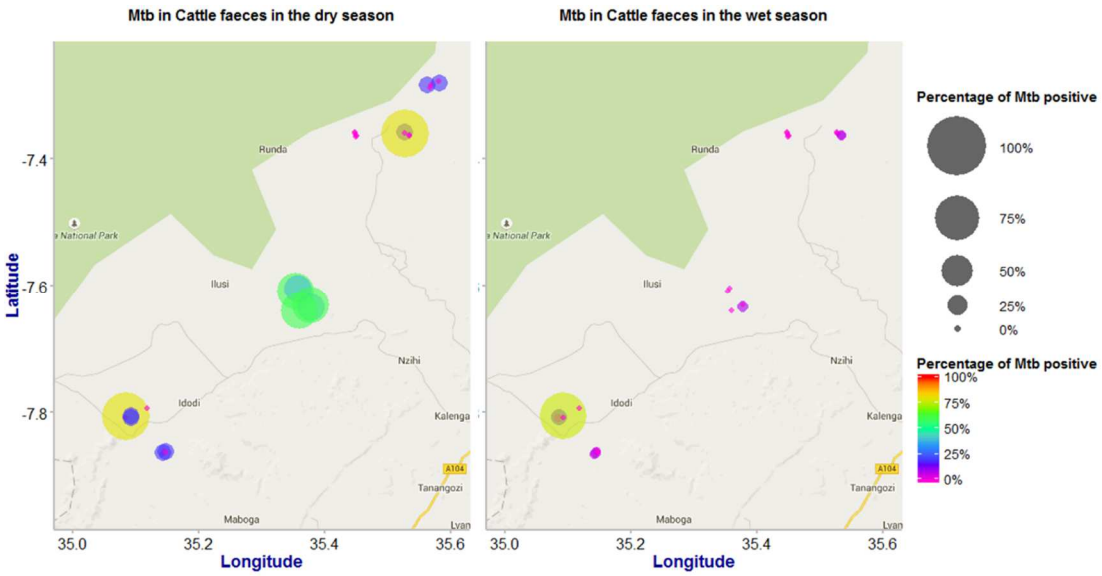


Figure 4.26 Mtb incidence in cattle faeces in the six villages in each season. No Mtb detection via goat faeces in both seasons.

There were very few Mtb positives in cattle boma soil with only one homestead in village 1 again in the wet season but none in the dry season (Figure 4.27). Mtb positives were under limit of detection via goat boma soil in both seasons.

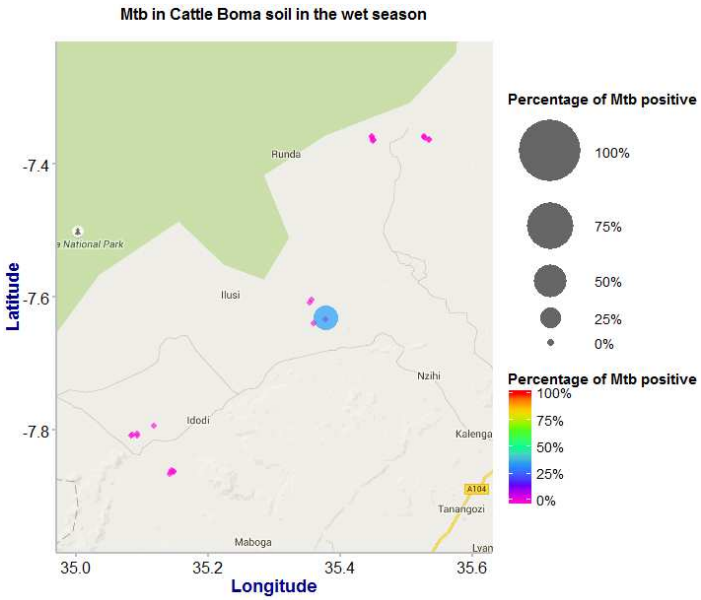


Figure 4.27 Mtb incidence in cattle boma soil in the six villages in the wet season. No Mtb detection in soil in the dry season

Mtb was shed positives were detected in household dust in both seasons with more than one households in village 5 (Figure 4.28).

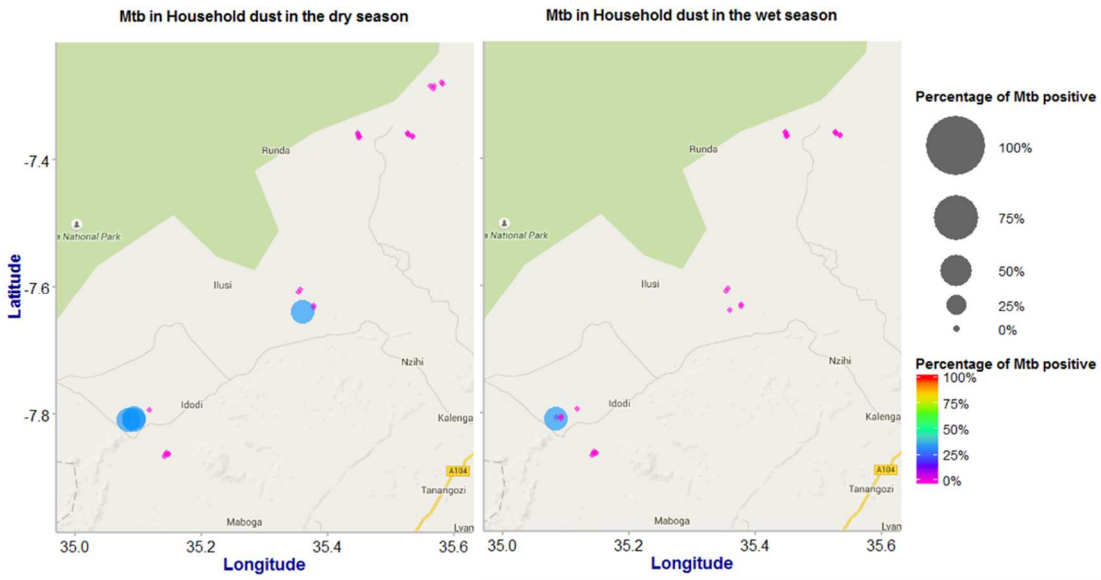


Figure 4.28 Mtb incidence in household dust in the six villages in each season.

The only one Mtb positive sediment is from Ikwavila (IKW) river site close to village 6 with Mtb shedding data via cattle faecal samples (Figure 29).

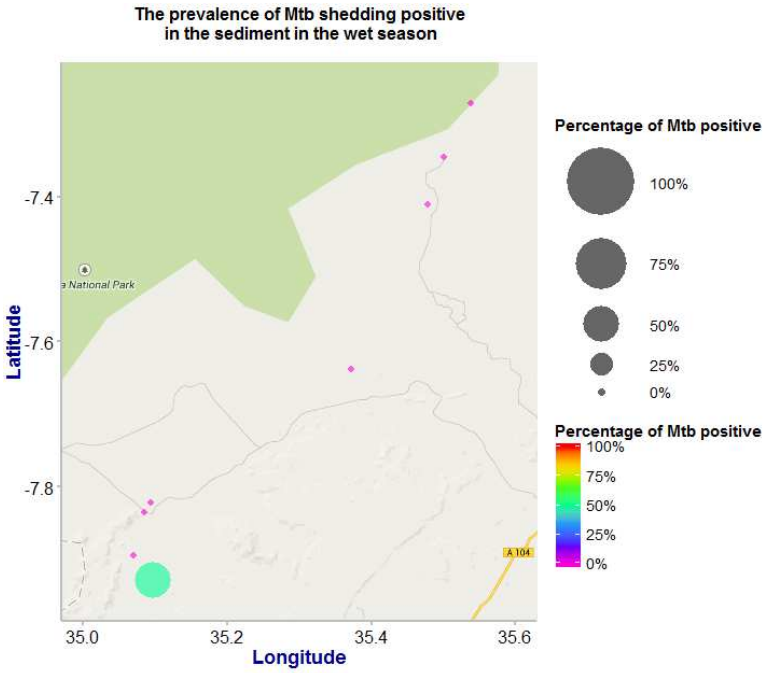


Figure 4.29 Mtb incidence in sediment samples in the eight river sampling sites in the wet season. No Mtb positives in sediment in the dry season.

4.4. Discussion

Several studies regarding bTB and TB in Tanzania indicated cross species transmission occurred within livestock, wildlife and human potentially via environmental reservoirs (Katale, Mbugi et al. 2012, Mbugi, Katale et al. 2012, Katale, Mbugi et al. 2013, Roug, Perez et al. 2014). The usual recognised route of cross species spread of bTB from animals to human is via the aerosol route or indirectly through raw milk consumption (Katale, Mbugi et al. 2012). Roug, *et al.* reported that unsterilised milk increased the potential risk with human exposure to the bTB if the infected cattle cannot be culled from cattle herds (Roug, Perez et al. 2014). Other SGM members such as MAP and *M. avium* were detected in soil and dust particles around infected

livestock and this could cause cross contamination from livestock to human (Eisenberg, Nielen et al. 2010, Kaevska, Lvoncik et al. 2014, Kolb, Hillemann et al. 2014). Very similar *M. avium* strains were reported in humans and their livestock which shared the same environment (Kolb, Hillemann et al. 2014). These studies all indicated the environment could play a role in transmission between man and livestock.

In the current study, prevalence data revealed that these two *Mycobacterium* species are ubiquitous and highly abundant in the livestock faeces, especially from cattle. In addition bioclimatic and environmental factor affecting moisture in soil also implied a critical role in both faeces and soil to explain the influence of moisture on survival of Mtb and Mb cells. One of the most striking features of the data is that faecal shedding indicates extensive infection as is often seen in wildlife where the disease is left unchecked (Hayley King 2015). This is evident for both cattle and goats but the interesting aspect relates to hosts tropism which for livestock relates to the prevalence of Mb in animals and Mtb in humans. Several herds of cattle appear to be shedding both Mb and Mtb, suggesting animals carrying one or other of these species. Detailed checks of preliminary data showed that several samples contained both pathogens which is very unusual. In addition village 1 is part of the NIH case positive household collection so humans there are infected with TB. Also in this village three samples had both pathogens Mtb and Mb within the faeces but no other samples indicated dual shedding across all positives for cattle or goat faeces. A report concerning cattle in Ethiopia discovered Mtb infection in cattle and attributed this to the habit of spitting chewed tobacco into the mouths of cattle, the practice was done

to reduce parasite load (Ameni, Vordermeier et al. 2011). Mtb is rarely reported in intensively reared cattle or in developed countries but is more common in developing area such as Sudan with 6.2 % (Sulieman and Hamid 2002), Algeria with 7.4 % (Boulahbal, Benelmouffok et al. 1978) and Ethiopia (Berg, Firdessa et al. 2009). The study of cattle receiving chewed tobacco in Selalle, Ethiopia indicated that 27 % of the isolates were Mtb (Ameni, Vordermeier et al. 2011).

Mb and Mtb were found more frequently in the dry season compared to the boma soil which might be expected given the high temperature and light irradiation effects reducing survival of both species in soil rather than in faeces which was sampled fresh. In agreement with the previous study in Ethiopia (Khera 2012) the water source was implicated in dissemination as villages 1 and 6 and 6 had the highest prevalence of Mb and both used water supplies from Ikonongo (IKO), Ikwavila (IKW) and Makife (NAK) (counts of 10^4 Mb). It was implied in villages 1 and 6 that the livestock shared the same environmental reservoir and water source which provided a potential risk factor for spreading disease within the homesteads and between cattle and goats especially as the cattle are regarded as the major reservoir for Mb transmission. Contamination of water supplies by cattle has been implicated in the spread of bTB in Africa (Humblet, Boschiroli et al. 2009).

The pathogens were able to survive in boma soil if high levels of cells were shed via faeces as indicated in villages 1 and 6 but only in the wet season.

The abundance of Mb and Mtb species estimated varied in environmental samples from 10^2 to 10^6 genome equivalents g^{-1} in solid samples and 10^2 to 10^4 gene copies $filter^{-1}$ in water samples. The survival of Mtb in faeces may be due to a number of

factors including dormancy, the thick waxy cell wall offering desiccation resistance (Gengenbacher and Kaufmann 2012) and possibly the richness of organic and inorganic material contained in livestock faeces which can improve the survival and proliferation of these two species (van Vliet, Reijs et al. 2007). In addition the maximum limit of bacterial cells extracted from soil using FastDNA Kit was 4 x 10^9 gene copies g^{-1}, which was proven in chapter 3. Therefore this result suggested that Mtb and Mb species was below 0.1 % in the whole microbial community of faeces and soil samples. A previous study even indicated the abundance of total micro-organism in water is 6 x 10^7 cells per ml^{-1} (Torsvik, Ovreas et al. 2002).

In conclusion, this study has revealed the significant reservoir of Mb in both cattle and goats in Tanzanian herds, coupled with the significant observation that Mtb appears to be shed by cattle. Systemic infections in animals with bTB can occur (Katale, Mbugi et al. 2012) but are rare in Mtb infections. Work within our research group has proven that faecal shedding is attributed to swallowing of sputum rather than lesions within the gut (tracheal shedding) (Hayley King 2015, King, Murphy et al. 2015). Environmental exposure reduced pathogen survival but the effect was less pronounced during the wet season. The local water supplies were implicated in the spread of disease but further work is required to establish transmission routes using strain typing and whole genome SNP analysis to prove sources of strains.

Chapter 5 Diversity analysis of SGM in Tanzanian pastoralist communities using 16S rRNA amplicon sequencing

5.1. Introduction

It was well proven that the qPCR molecular approach with specific primers provided successful identification of target cells from environmental samples used for screening in this study (Chapter 4). It was evident from isolation work conducted at the various abattoirs that non-tuberculous mycobacteria were being isolated at a high frequency from lesions in lungs and other tissues (Table 2.4). It was therefore decided to obtain an in-depth study of the diversity within the samples taken to investigate prevalence of other putative mycobacterial pathogens present in the environment. This species variation can be referred to as genetic variation on alpha or beta diversity from the environment to provide potential correlations with metadata such as human activities, livestock husbandry or wildlife management (Larsson 2001).

The 16S rRNA gene was used for the establishment of taxonomic diversity due to its unique, highly conserved sequence in eubacteria. The SGM 16S rRNA V3 region target primer APTK was developed to screen SGM species from the environment allowing specific amplification of only a subgroup within the *Mycobacterium* genus which contains the long 18 helix unique to SGM (Pontiroli, Khera et al. 2013). All SGM have this insertion with rare exceptions and the important pathogen species complexes are not excluded (Harmsen, Dostal et al. 2003). The modification was applied for 454 pyrosequencing (Pontiroli, Khera et al. 2013) and Miseq analysis. The specificity and

sensitivity of APTK primer for detection of SGM was well proven and applied on 454 pyrosequencing in bio-diversity analysis (Khera 2012, Pontiroli, Khera et al. 2013).

High throughput pyrosequencing offered an opportunity to investigate in depth the diversity of mycobacteria in soil and water in previous studies (Khera 2012, Pontiroli, Khera et al. 2013). This sequence shotgun approach was developed based on detection of pyrophosphate release on nucleotide incorporation during sequence synthesis (Ronaghi, Uhlen et al. 1998). The order of nucleotides was recorded while the signal of excitation state of luciferase fluorescence was detected by monitor. This advantage was contributed to generate 400 Mb within 10 hours and provided new dimension for genome sequencing and metagenomics analysis (Ronaghi, Uhlen et al. 1998).

The sequencing analysis platform, QIIME, was applied after 454 pyrosequencing to develop a complete survey of the whole bacterial community distribution in an ecosystem. QIIME was used to bin sequence reads as clusters of OTU grouped at a specified level of similarity (97 % similarity as default). For alpha diversity the candidate sequence was identified based on reference to online databases such as BLAST, RDP and Greengenes. The phylogenetic tree and diversity analysis were then introduced to estimate the similarity of each OTU cluster and correlation among samples.

The bio-diversity of OTU from environmental samples was presented by alpha and beta diversity which was introduced by *Whittaker* (Whittaker 1960). The definition of alpha diversity was slightly controversial but the main idea was to exhibit the mean of species diversity in different habitats or sample collection subunits within land

scales. The beta diversity delineated the difference or similarity of species community composition among local land scales. These two diversity measurements were compared at sequence level to analyse all SGM sequences using QIIME.

The Shannon index measure was introduced to describe the species richness and evenness within the samples in alpha diversity analysis. The Principle Component Analysis (PCA) and Principle Coordinate Analysis (PCoA) was an ordination tool for exploratory data analysis in beta diversity. The clusters of different sample sets in these two analysis figures were plotted in two or three dimensions on their corresponding axes (Hotelling. 1933). The similarity or difference of bacterial species communities was estimated as the proportion of overlap region between diverse sets. The PCA was a metagenomics analysis tool for large variance but it is still require to normalise each species sum to give abundance of 1 to draw of this figure. In comparison to PCA, PCoA conducted an ordination or user defined dissimilarity measure using Bray-Curtis index and was not limited within 0 to 1 because this index was used in the ordination rather than species abundance. The relationship of metagenomics profile was more flexible and appropriated using Bray-Curtis distance in PCoA compared to PCA, and supports at the high level taxonomic analysis and diversity analysis (V. Kunin 2008).

Another plot used in this study was NMDS plot based on Bray-Curtis ranking distance matrix with multivariate analysis method. The NMDS plot focuses on relative position between samples types which are in different habitat or climatic conditions into two or three dimensional ordination space (Kruskal. 1964).

The environmental diversity analysis relied on the two components, richness and evenness. The richness referred to the numbers of different species obtained in an ecological system and evenness indicated the similarity in numbers between two groups. The species richness was correlated with species distribution between group types and spatial factors such as the similarity of two bacterial communities compared between two different regions or sample set. It is crucial for epidemiology that spatial and temporal analysis can be done to determine pathogen distribution with disease and geography.

5.2. Aims

1. To investigate the abundance and distribution of SGM in environmental samples using 16S rRNA amplicon analysis.

2. To compare different methods in QIIME for diversity analysis and select appropriated approach for environmental screening

3. To make the correlation between human, livestock and environment for evaluation of pathogen reservoirs.

5.3. Results

5.3.1. 16S rRNA amplicon Pyrosequencing and OTU abundance distribution

The 454 pyrosequencing was performed on total community DNA from 58 pilot environmental samples using SGM APTK primers sets. This sequencing run produces a total 547,801 sequences. The distribution of sequence length (Figure 5.1) indicated

that the majority of sequence length was 429 bp at 36.2 % after trimming. The population of 70 % sequence length ranged between 420 to 430 bp. Sequences were retained if the amplicon length was between 420 ~ 430 bp depended on the expected amplicon length.

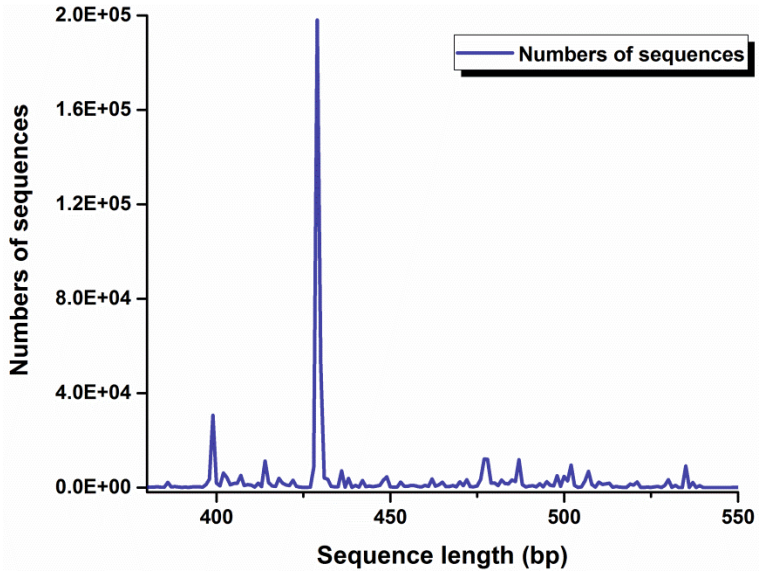


Figure 5.1 The length of sequences used for SGM analysis

The QIIME pipeline was used for analysis of alpha and beta diversity. The sequence data were clustered using OTU similarity approach based on user defined similarity threshold. The different sequence similarity threshold was used to cluster form OTU to identify species, genus, family and even order (Figure 5.2). When clustered at 97 % similarity 72.4 % of sequence were assigned to the *Mycobacterium* genus whereas 79 % are assigned to this genus when clustering at 80 % similarity. In addition the proportion of unassigned group was increased from 12.9 % to 13.2 % by decreasing the sequence similarity. The 3 % dissimilarity threshold for taxonomic assignment is still controversial and a 98.7 % sequence similarity threshold was recommended for species level recently, especially on species identification or diversity analysis using

16S rRNA (Konstantinidis, Ramette et al. 2006, Stackebrandt 2006, Webster, Taylor et al. 2010).

Another issue associated with taxonomic assignment was the particular proportion named non-BLAST hit which referred to no sequence reference match to curated database chosen. The occurrence of non-BLAST hit was because most of the curated database still relies on isolation of micro-organisms. For example, Bergey's Manual provided > 5,000 species 16S rRNA (Garrity 2004) and Standing in Nomenclature database collected > 12,000 species in total (Euzeby 1997) representing only a part of estimated microbial diversity in environment (Eren, Maignien et al. 2013).

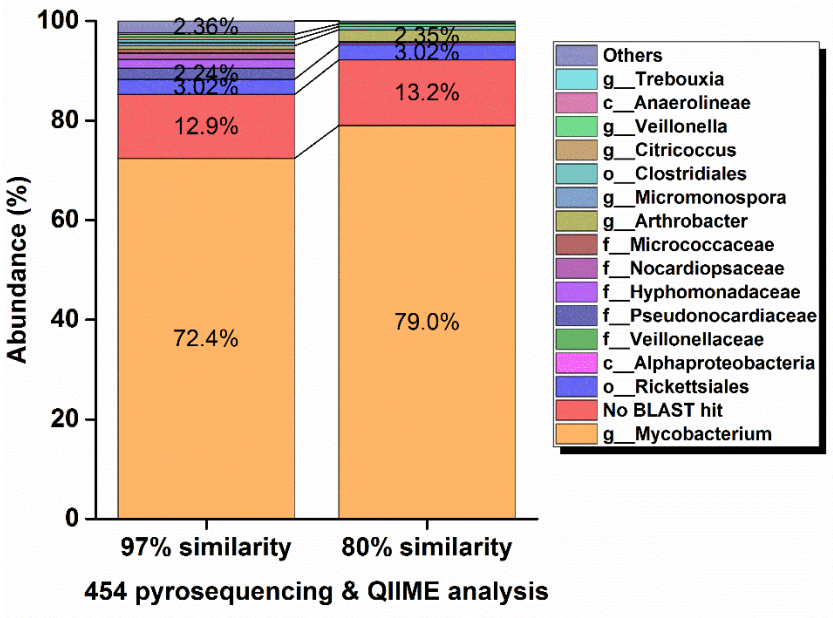


Figure 5.2 The proportion of diverse bacterial genus, family, order and no BLAST hit based on coefficient of similarity.

This comparison was illustrated and highlighted the *Mycobacterium* and non-BLAST hit which refer to no reference matched in BLAST database, between samples (Figure

5.3). These proportions of unassigned taxonomic units are observed in water samples at 38.0 % of whole water samples sequence. The other sample types in contrast only ranged up to 10 % of sequences with non-BLAST hits. The bacterial distribution between household dust and sediment samples was similar (Figure 5.3).

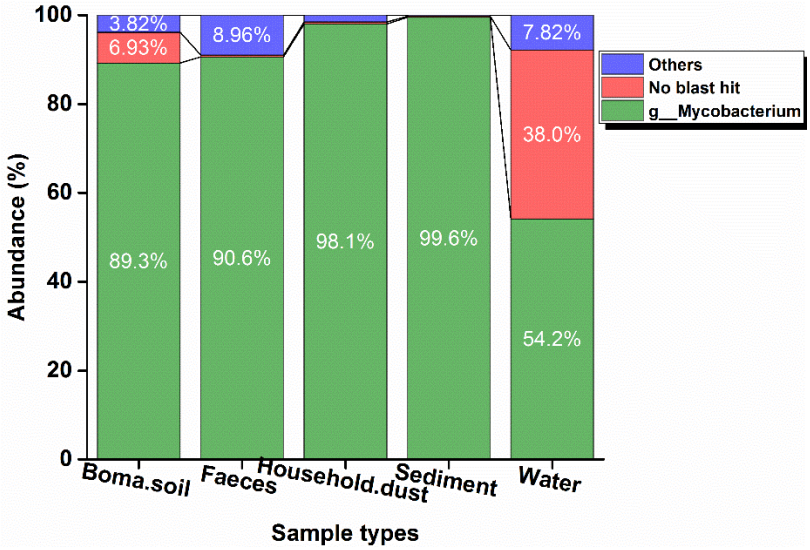


Figure 5.3 Different proportion of taxonomic unit assigned in diverse samples sets.

For the purpose of further analysis cattle faecal and cattle boma soil samples are analysed together and called “cattle-related samples”, as are goat faecal and boma samples are named “goat-related samples”. The same purposes are provided to household dust as “human-related samples” and water and sediment samples are “water-related samples”. Different samples sets were associated with different organisms, for instance, cattle faecal samples were distinct form goat faecal samples. Human-related samples contained a higher proportion of mycobacteria than animal or environmental samples (Figure 5.4), with the lowest abundance of *Mycobacterium* genus occurring in goat related samples where they were only four-fifths of the community.

There were similar bacterial distributions among the three sample collection areas comprising of PAWAGA division, the north area including 2nd,3rd and 4th villages, middle region only covering 1st village are and IDODI division, the south region with 5th and 6th villages (Figure 5.5). 92.5 % of sequences were part of the *Mycobacterium* genus in middle area compared to other two regions where they were 92.9 % and 88.6 % in PAWAGA and IDODI division respectively. It was worth noting that of the proportion of non-BLAST hits was twofold higher in PAWAGA than in middle area and IDODI.

Prior to 454 pyrosequencing samples were screened for SGM positivity using PCR and gel electrophoresis and even those found to be negative by gel electrophoresis did contain mycobacteria when their communities were examined by pyrosequencing. The proportions of *Mycobacterium* genus was similar in gel electrophoresis samples with test positives containing of 91.2 % *Mycobacterium* genus compared to 89.7 % in SGM test negative. It was implied the lack of accurate identification of concealed bacterial species using gel electrophoresis occurred, especially for environmental screening and diversity analysis.

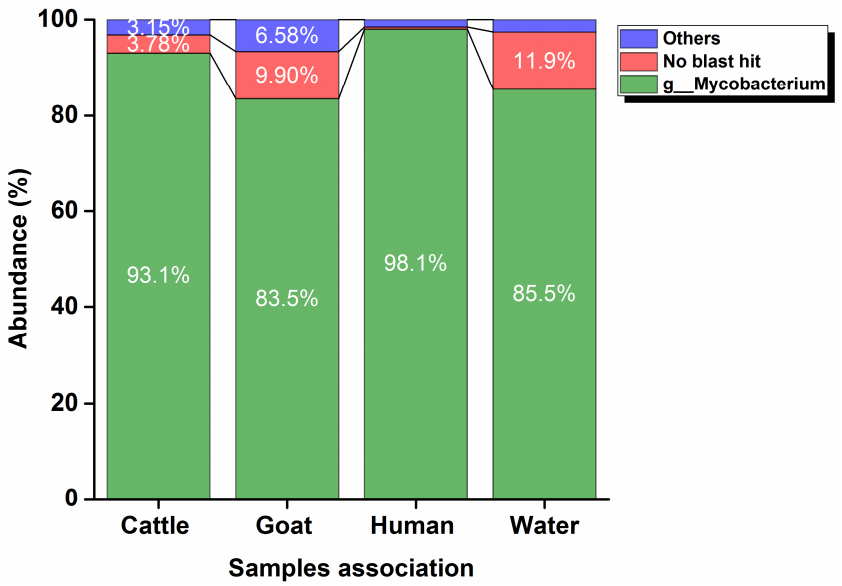


Figure 5.4 Different proportion of taxonomic unit assigned in diverse sample association.

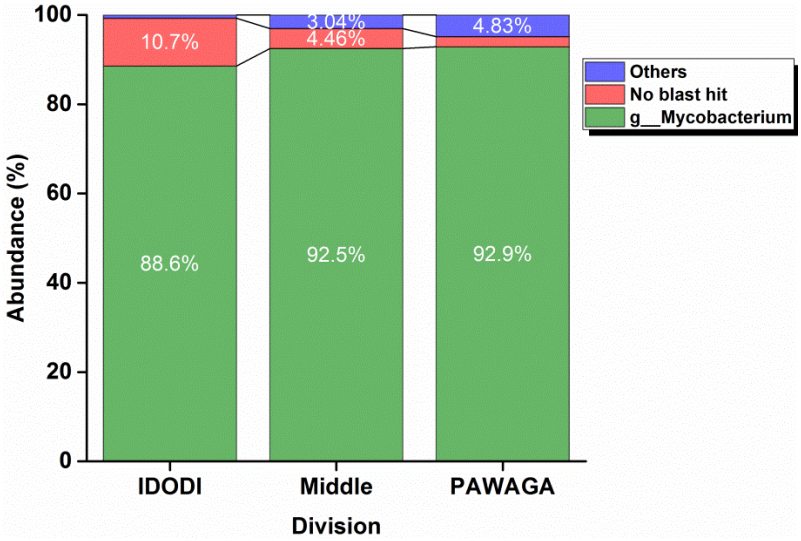


Figure 5.5 Different proportion of taxonomic unit assigned in diverse sample region.

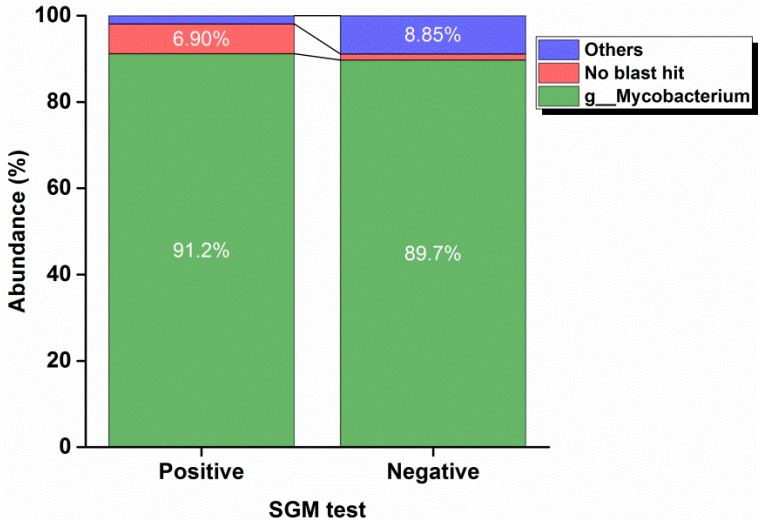


Figure 5.6 Different proportion of taxonomic unit assigned in two SGM test sets.

5.3.2. The alpha and beta diversity of SGM species

The Shannon diversity index was as a measure of alpha diversity for comparison of both species richness and abundance from diverse habitats or samples sets. The Shannon index was affected by both species richness and abundance, was different from other three indexes in alpha diversity description, such as Simpson index forces on species abundance only, Chao1 metric estimate the species richness and the Observed Species metric simply counted the unique OTU found in samples (Kuczynski, Stombaugh et al. 2012).

5.3.2.1. Alpha diversity

The rarefaction curve illustrates that the random collection of sequence reads was estimated from each sample with specific depth (number of sequences) (Figure 5.7). All samples can reach the plateau at less than 500 sequences so the depth of 300 sequence was randomly selected for all sample categories to compare the richness of samples sets. The comparison between different sample types was depicted in

Figure 5.8 and no significant difference between each pair using statistical analysis t-test and P-value (Table 5.1). It was indicated that the household dust samples was the lowest species abundance contained among these five sets.

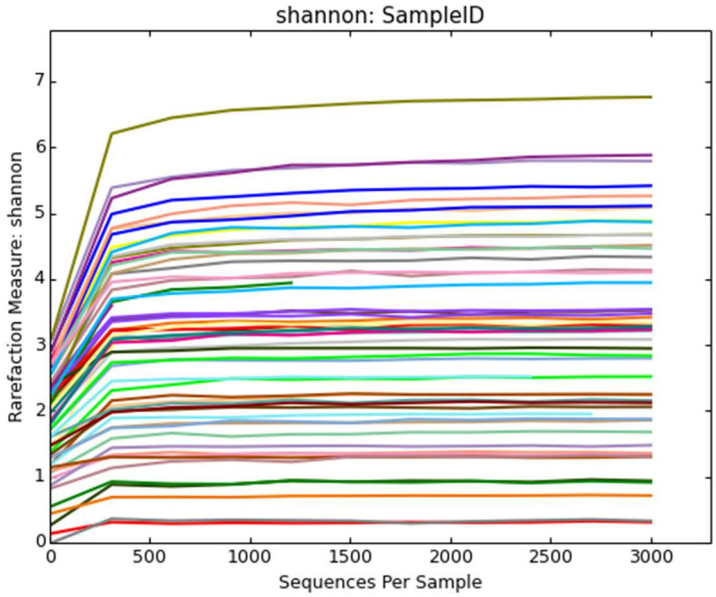


Figure 5.7 Rarefaction curve for sequence of all samples using APTK primer with Shannon measure index. Different samples were highlighted in different colours.

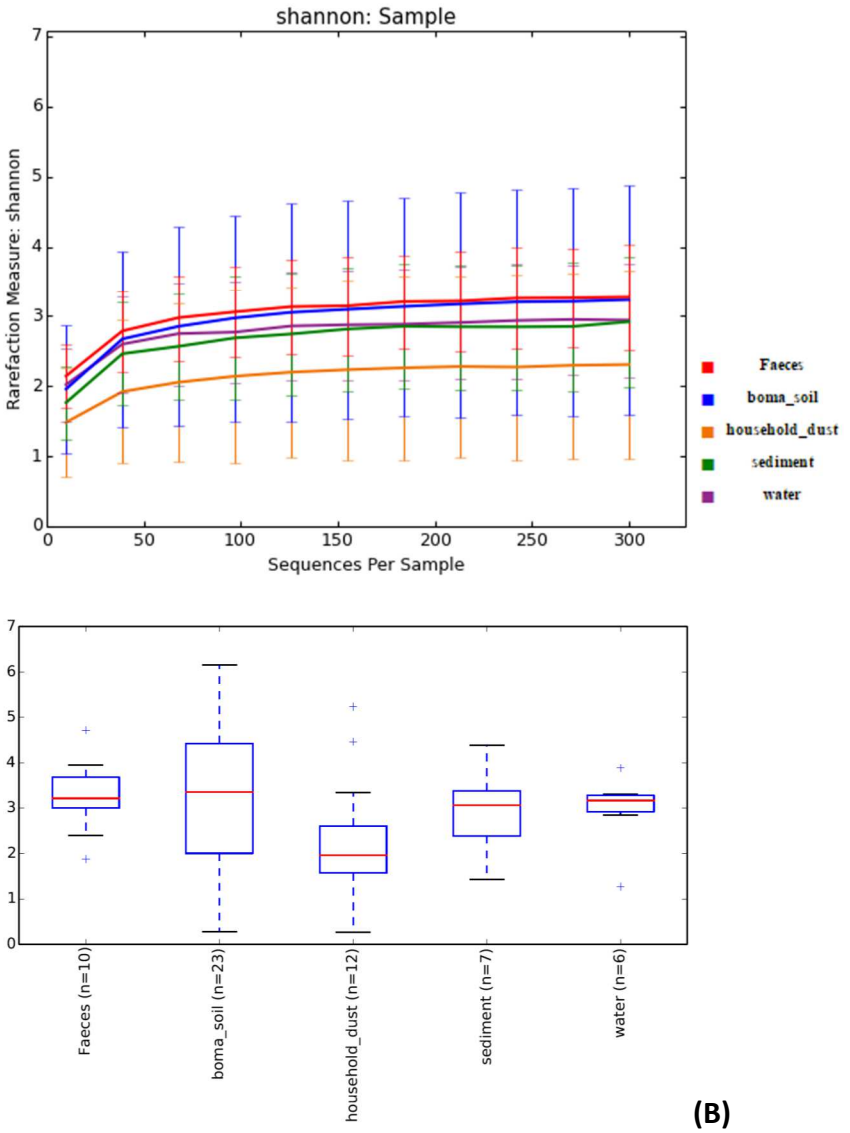


Figure 5.8 Rarefaction curve (A) and range (B) for sequence of all samples clustered into five samples sets using SGM 16S rRNA primer with Shannon measure index.

Table 5.1 Comparison between each pairs of samples types using statistical analysis t test and P-value. ***Indicates significant difference (P-value < 0.001) using t-test and P-value statistical analysis. F: faeces, HH: household dust, BS: boma soil, S:sediment and W:water.

Group1	Group2	Group1 mean	Group1 std	Group2 mean	Group2 std	t stat	P-value
W	S	2.940045	0.808099	2.918614	0.942444	0.04013	1
HH	F	2.307116	1.339185	3.270706	0.753195	-1.92997	0.68
F	S	3.270706	0.753195	2.918614	0.942444	0.802467	1
BS	S	3.232964	1.64414	2.918614	0.942444	0.465957	1
W	F	2.940045	0.808099	3.270706	0.753195	-0.77362	1
HH	BS	2.307116	1.339185	3.232964	1.64414	-1.63256	1
HH	S	2.307116	1.339185	2.918614	0.942444	-1.00656	1
BS	F	3.232964	1.64414	3.270706	0.753195	-0.06735	1
HH	W	2.307116	1.339185	2.940045	0.808099	-1.00391	1
W	BS	2.940045	0.808099	3.232964	1.64414	-0.40841	1

The significant difference was observed between cattle related samples and human related samples (Figure 5.9 and Table 5.2). It was estimated that the cattle related samples had the highest species richness and abundance while human related samples had the lowest. It was worth noting that the similar diversity values within human, goat and water related samples may be as a result of high similarity on species richness and abundance within these three samples sets.

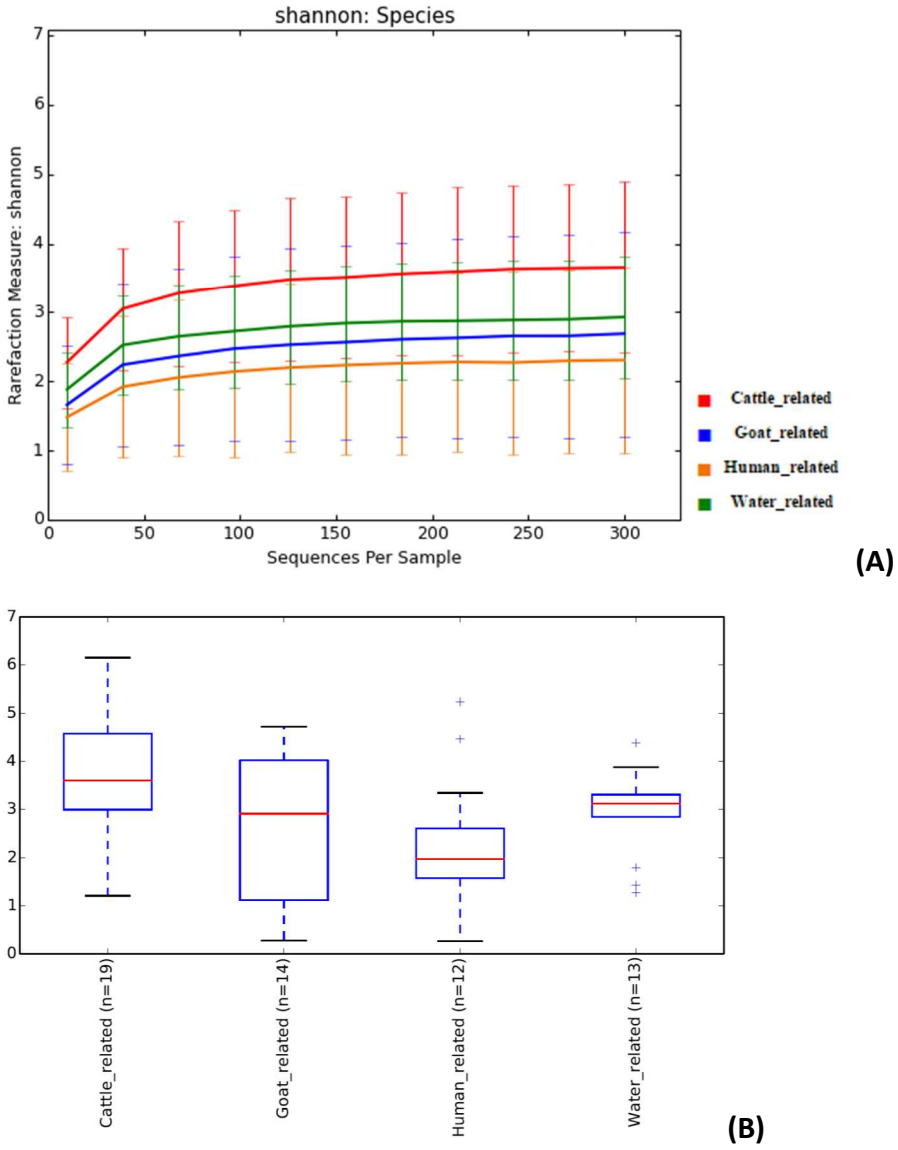


Figure 5.9 Rarefaction curve (A) and range (B) for sequence of all samples clustered into four samples sets using SGM 16S rRNA primer with Shannon measure index.

Table 5.2 Comparison between each pairs of samples types using statistical analysis t test and P-value. *Indicates significant difference (P-value < 0.05) using t-test and P-value statistical analysis. C: cattle, H: human, G: goat and W: water.

Group1	Group2	Group1 mean	Group1 std	Group2 mean	Group2 std	t stat	P-value
H_related	W_related	2.307116	1.339185	2.928505	0.883047	-1.32305	1
G_related	H_related	2.687962	1.484095	2.307116	1.339185	0.655446	1
H_related	C_related	2.307116	1.339185	3.654408	1.245881	-2.7549	0.042*
C_related	W_related	3.654408	1.245881	2.928505	0.883047	1.75471	0.492
G_related	C_related	2.687962	1.484095	3.654408	1.245881	-1.96691	0.294
G_related	W_related	2.687962	1.484095	2.928505	0.883047	-0.48783	1

The IDODI region had the least species richness and abundance compared to the other two regions as demonstrated by Shannon index (Figure 5.10 and Table 5.3). In addition the IDODI division is significantly different to other two regions as evaluation of P-value (T-test, $p < 0.05$). Another estimation was to compare SGM positive and negative samples as determined during the pre-tests using gel electrophoresis with APTK primer as the same primer using in 454 pyrosequencing (Figure 5.11 and Table 5.4). Similarities in two pre-test types are observed in the Shannon estimation plot (Figure 5.11). There was no significant difference between two pre-test types as identified as SGM positive and negative in Shannon measure estimation (T-test, $p > 0.05$).

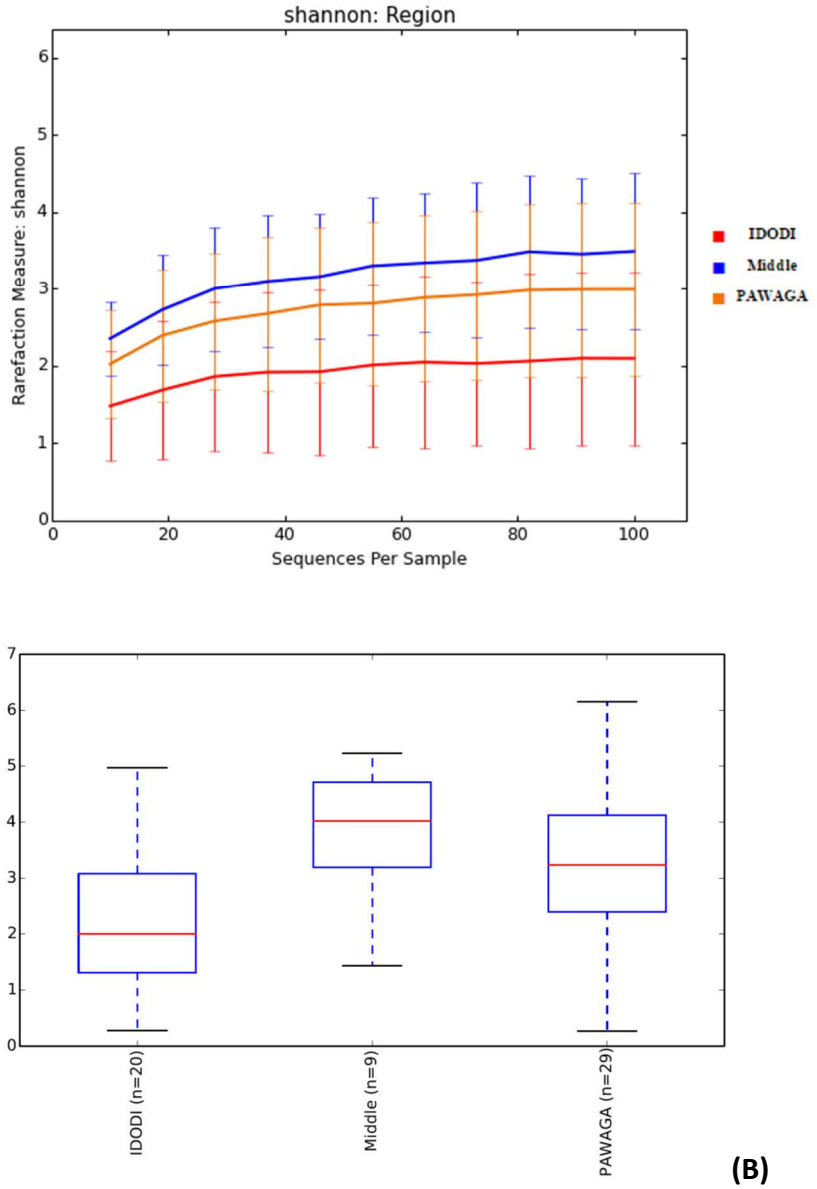


Figure 5.10 Rarefaction curve (A) and range (B) for sequence of all samples clustered into three samples sets using SGM 16S rRNA primer with Shannon measure index.

Table 5.3 Comparison between each pairs of samples types using statistical analysis t test and P-value. *Indicates significant difference (P-value < 0.05) using t-test and P-value statistical analysis.

Group1	Group2	Group1 mean	Group1 std	Group2 mean	Group2 std	t stat	P-value
Middle	IDODI	3.766429	1.140127	2.257602	1.248671	2.982758	0.027*
IDODI	PAWAGA	2.257602	1.248671	3.233493	1.266814	-2.6109	0.048*
Middle	PAWAGA	3.766429	1.140127	3.233493	1.266814	1.098118	0.81

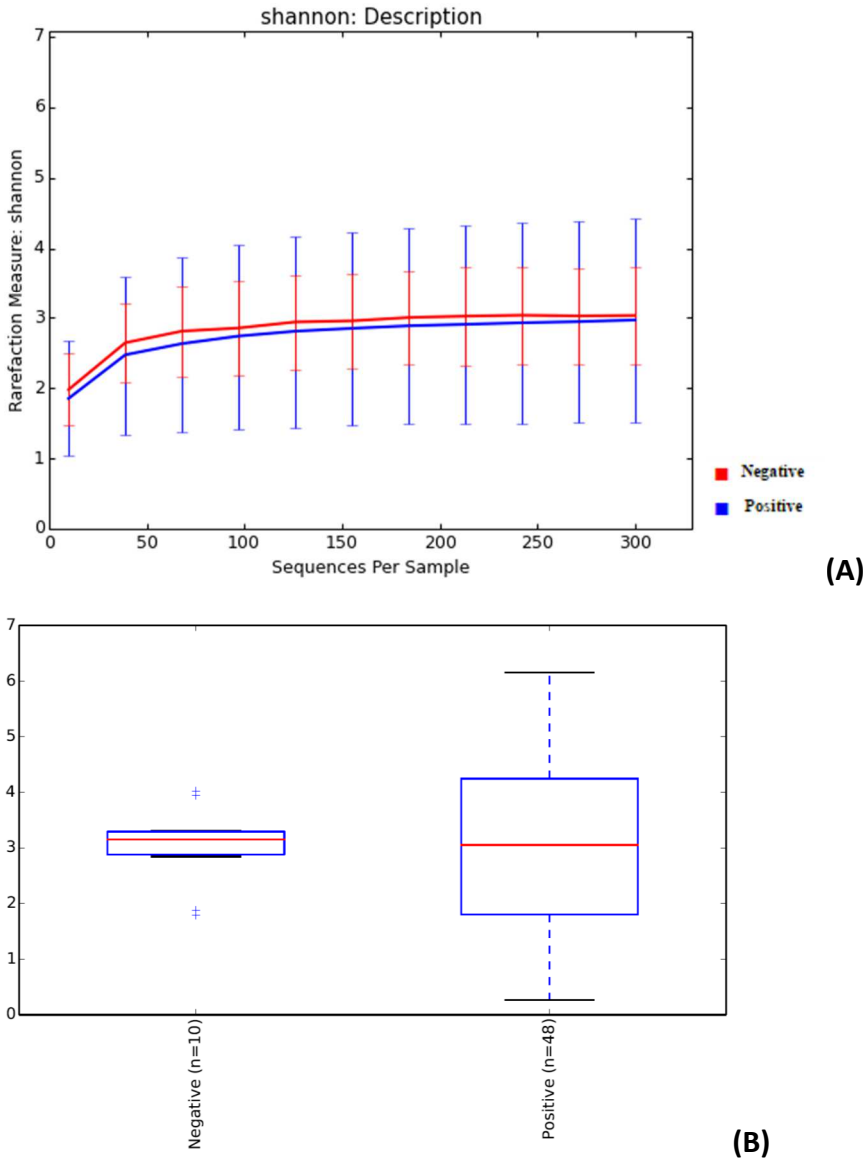


Figure 5.11 Rarefaction curve (A) and range (B) for sequence of all samples clustered into two samples sets using SGM 16S rRNA primer with Shannon measure index.

Table 5.4 Comparison between each pairs of samples types using statistical analysis t test and P-value. *Indicates significant difference (P-value < 0.05) using t-test and P-value statistical analysis.

Group1	Group2	Group1 mean	Group1 std	Group2 mean	Group2 std	t stat	P-value
Positive	Negative	2.968564	1.460571	3.033013	0.696988	-0.13397	0.882

The alpha diversity demonstrated that species richness and abundance was influenced by sample types and regions. It was noticed that the bacterial communities in diverse environmental reservoirs are variable due to animal or

human shedding and interaction. The beta-diversity later provided evidence regarding to bacterial population distribution and similarity based on shared OTU.

5.3.2.2. Beta diversity

The distance matrix based beta diversity measures combined with quantitative and qualitative approach was a very powerful tool to analysis differences and similarities among bacterial communities. The analysis now was based on both UniFrac Weighted and Unweighted analysis of beta diversity. The UniFrac measures included in beta diversity was based on phylogenetic information for comparison of environmental samples or two collections of sequences, for instance, 16S rRNA gene sequence from diverse microbial samples (Lozupone, Lladser et al. 2011). The Unweighted analysis is a qualitative test concerning the presence/absence of OTU whereas Weighted analysis is a quantitative test looking at the relative prevalence of OTU (Lozupone, Hamady et al. 2007, Khera 2012).

The *Mycobacterium* species community composition in different sample types were compared using PCoA plots, NMDS plots and ANOSIM statistical analysis on dissimilarity multiple variation. Overlapping regions of sample sets occurred in both Weighted and Unweighted analysis among household, sediment and water samples (Figure 5.12A & B) in both PCoA and NMDS plot. Most samples in the Weighted analysis cluster together in the upper-right region, which demonstrates the high similarity in relative abundance of OTU between samples sets (Figure 5.12A). Dispersion of samples in Unweighted analysis, however, compared to Weighted analysis provides strong correlation with the presence/absence of OTU in sample sets.

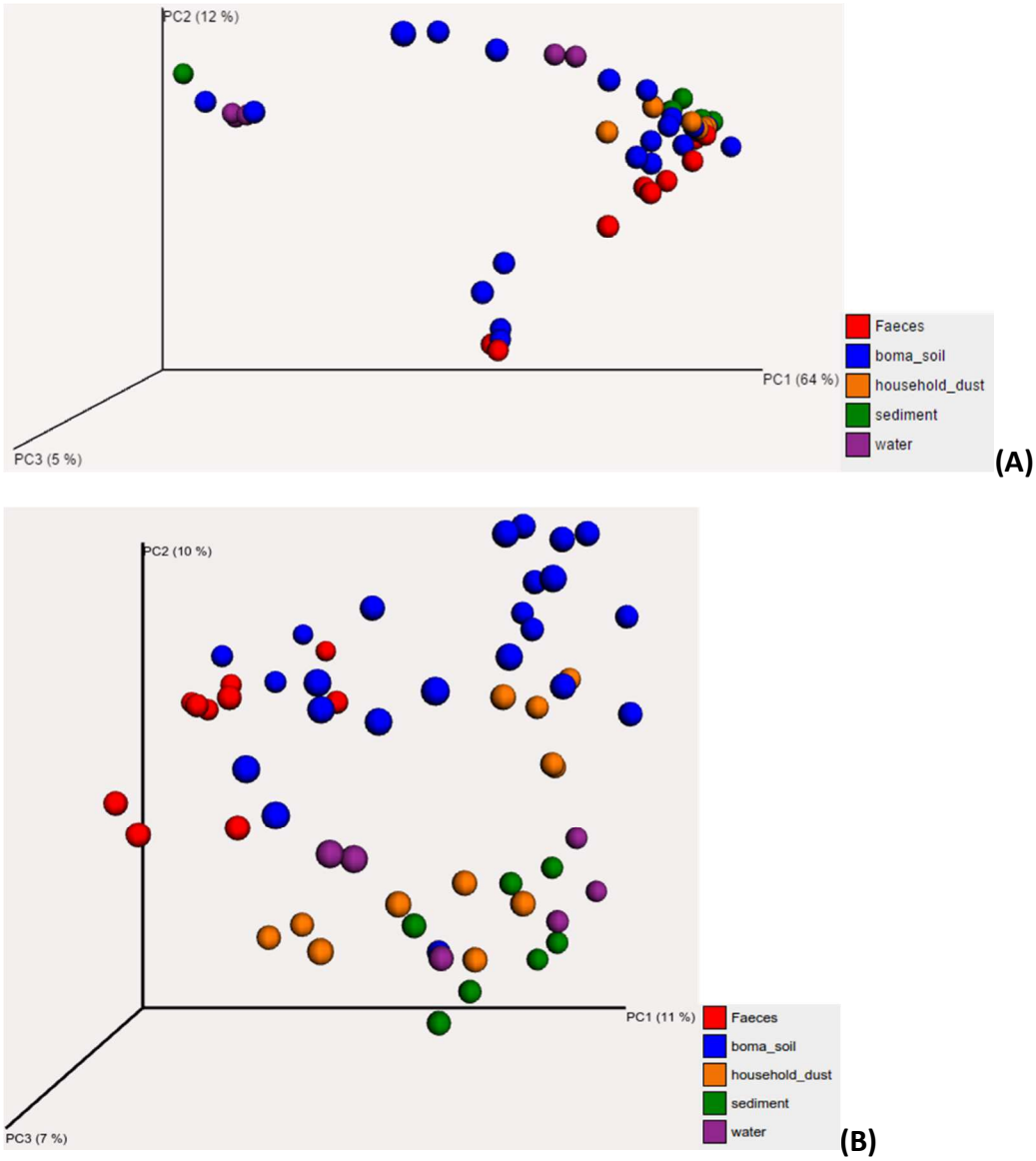
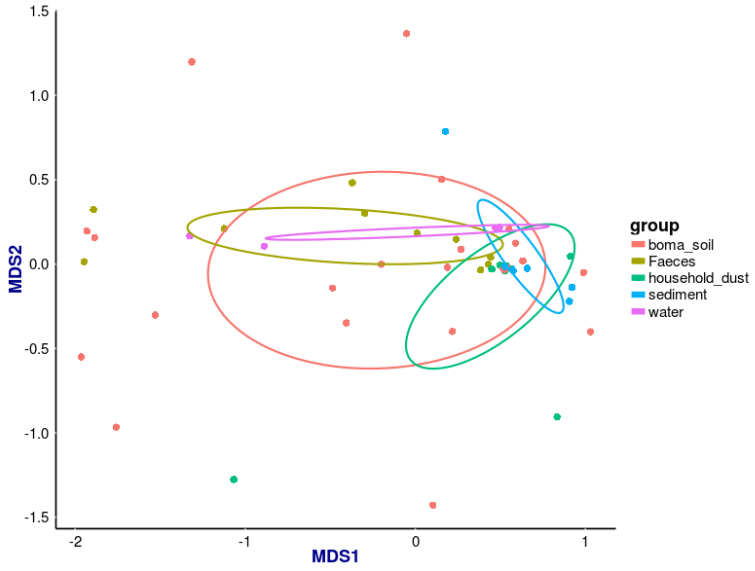


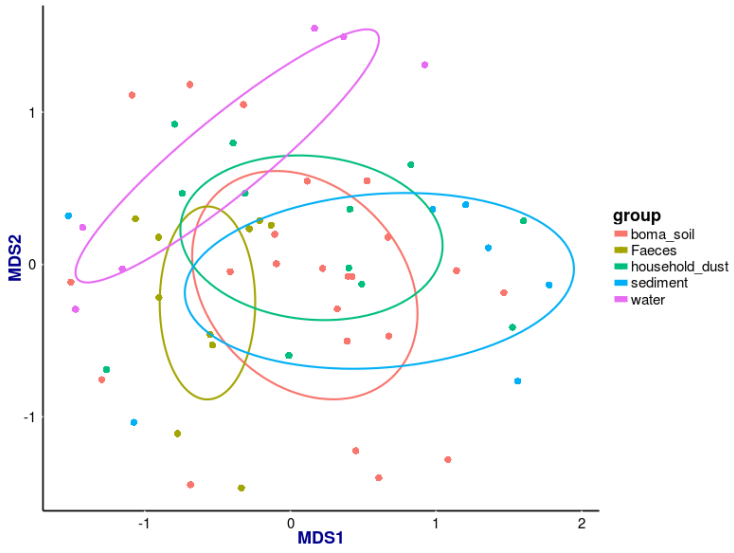
Figure 5.12 Comparison of environmental samples sets using **(A)** Weighted analysis **(B)** Unweighted analysis in PCoA plot.

Similar to the PCoA figure, all samples overlapped together in a region in the NMDS plot, which demonstrated again high similarity between samples types in Weighted analysis in relative positions (Figure 5.13A). In contrast to OTU abundance, water samples was highly separate and different compared to other four samples types in terms of presence/absence OTU (Figure 5.13B). In addition boma soil samples clustered with faecal and household dust sample separately in both Weighted and Unweighted analysis but latter two samples types were not overlapped together as

many as boma soil samples. The NMDs plot suggests that boma soil samples shared higher similarity of OTU abundance and presence/absence to faeces and household dust samples than faecal samples shared to household dust samples.



(A)



(B)

Figure 5.13 Comparison of environmental samples sets using **(A)** Weighted analysis **(B)** Unweighted analysis in NMDS plot.

The statistical analysis using the ANOSIM test to compare each pairwise samples was based on their OTU variable composition (Table 5.5). A classical t-test only performs on the pairwise group dispersion but ANOSIM analysis combined with a permutation test based on the classical t-test on pairwise group dispersion to the non-normal distribution. Several significant differences occurred in OTU abundance between sample types (Table 5.5A). In contrast to OTU abundance, all pairs differed In OTU presence/absence (Table 5.5B). The ANOSIM result provides the significance to prove the similarity and difference in depth to pairwise comparison of samples and a similar observation was found as PCoA plots shown between household dust, water and sediment samples.

Table 5.5 Comparison of environmental sample types. The samples are denoted by the abbreviation of samples: HH: household dust, W: water, S: sediment, BS: boma soil and F: faeces. The significant difference is represented by P-value. * ANOSIM test (P-value < 0.05), ** (P-value < 0.01) and *** (P-value < 0.001) in **(A)** Weighted analysis **(B)** Unweighted analysis.

Pairwise comparisons for QIIME analysis (P-value) in Weighted analysis (A)

	BS	F	HH	S	W
BS		0.696	0.763	0.586	0.003**
F	0.696		0.001***	0.067	0.002**
HH	0.763	0.001***		0.056	0.001***
S	0.586	0.067	0.056		0.014*
W	0.003**	0.002**	0.001***	0.014*	

Pairwise comparisons for QIIME analysis (P-value) in Unweighted analysis (B)

	BS	F	HH	S	W
BS		0.001***	0.004**	0.002**	0.001***
F	0.001***		0.001***	0.001***	0.001***
HH	0.004**	0.001***		0.001***	0.001***
S	0.002**	0.001***	0.001***		0.001***
W	0.001***	0.001***	0.001***	0.001***	

No segregation of samples sets is observed under Weighted analysis in both PCoA plot and NMDS plot (Figure 5.14A and 5.15A). The same situation occurred in Unweighted analysis in NMDS (Figure 5.15B) but not in the PCoA plots. The PCoA figure in the Unweighted analysis, which illustrates water related samples were constantly the most significant factor in the OTU presence/absence of shared OTU. However weak segregation was shown within cattle and goat related samples in both plots and ANOSIM statistical analysis proves the high similarity between these pair samples (Table 5.6). Nevertheless the statistical analysis indicates differences

regarding water related samples to other three datasets, showing significance only between water/human, but not for water/cattle and water/goat in Weighted analsis and highly significant in water related samples to other three datasets in Unweighted analysis (Table 5.6). It was thought that the OTU abundnce of shared microbial communities in water and animals rather than humans are highly similar because livestocks use these river as water source and share the same bacterial communities from water source between herds.

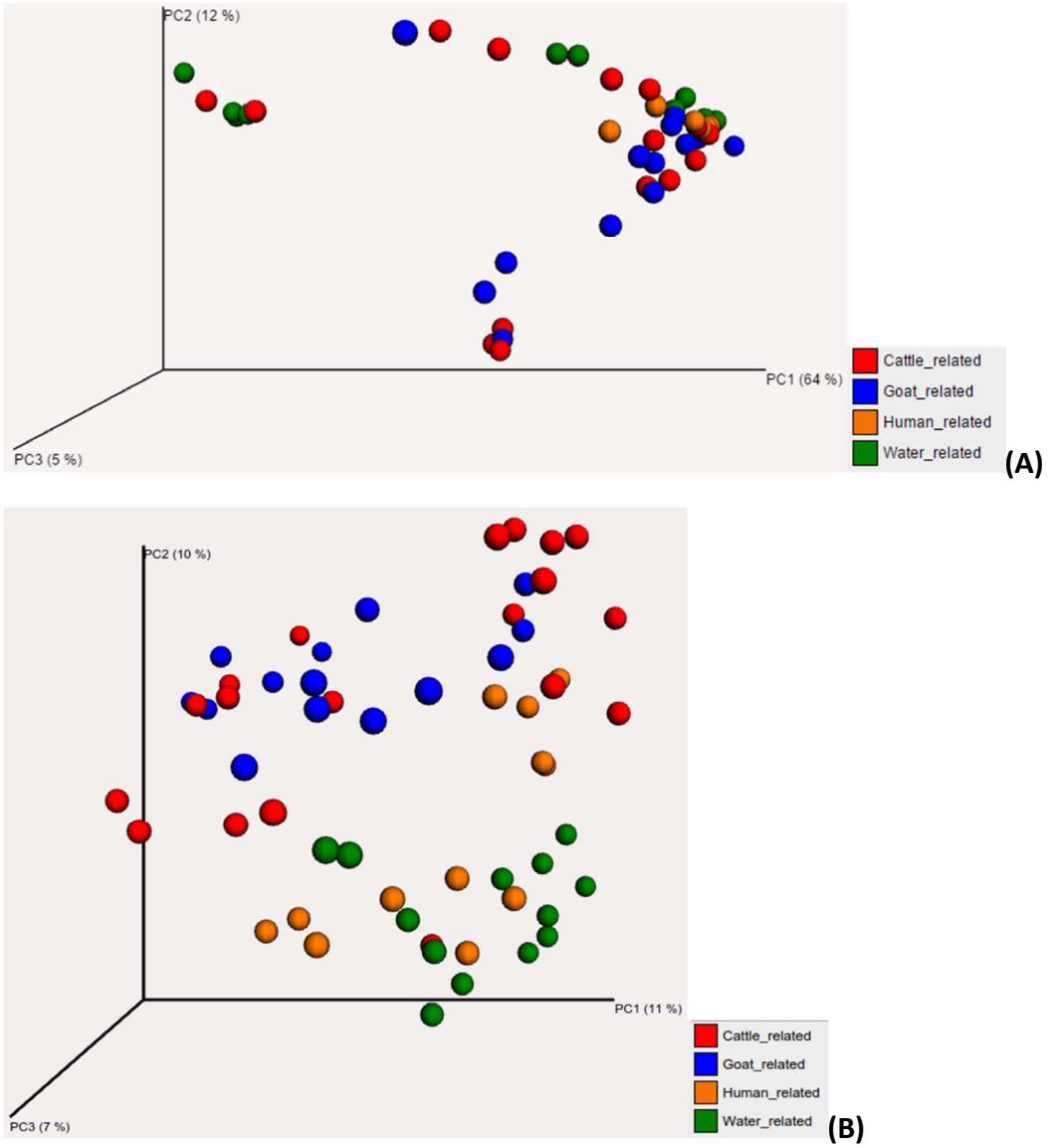
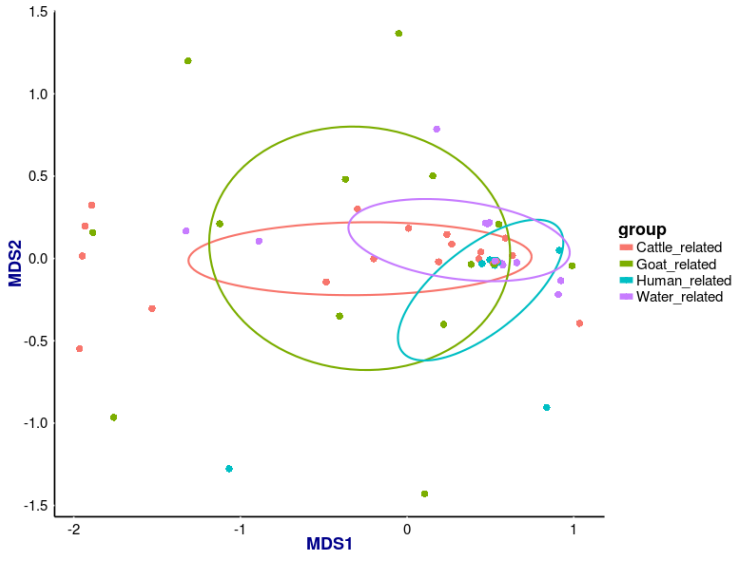
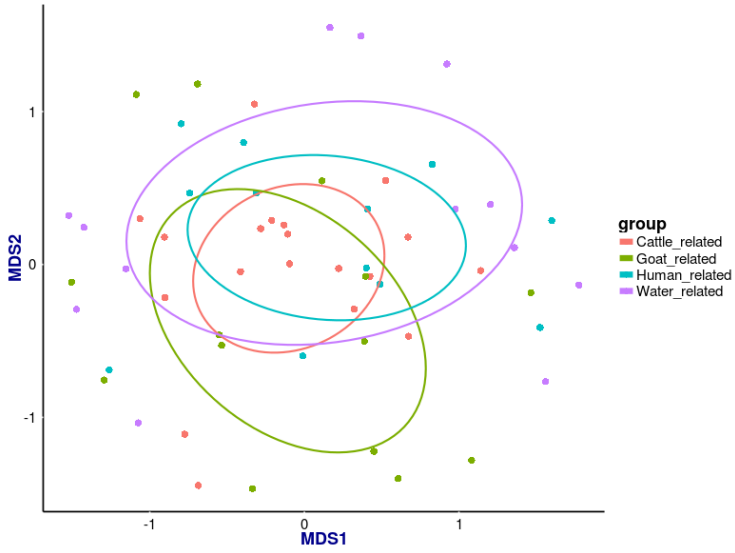


Figure 5.14 Comparison of environmental samples types suing **(A)** Weighted analysis **(B)** Unweighted analysis in PCoA plot.



(A)



(B)

Figure 5.15 Comparison of environmental samples types using **(A)** Weighted analysis **(B)** Unweighted analysis in NMDS plot.

Table 5.6 Comparison of environmental samples types. The samples are denoted by the abbreviation of samples: HH: household dust, C: cattle related samples, G: goat related samples, H: human related samples and W: water related samples. The significant difference is represented by P-value. * ANOSIM test (P-value < 0.05), ** (P-value < 0.01) *** (P-value < 0.001) in **(A)** Weighted analysis **(B)** Unweighted analysis.

Pairwise comparisons for QIIME analysis (P-value) in Weighted analysis (A)

	C	G	H	W
Cattle related		0.533	0.466	0.055
Goat related	0.533		0.013*	0.030*
Human related	0.466	0.013*		0.007**
Water related	0.055	0.030*	0.007**	

Pairwise comparisons for QIIME analysis (P-value) in Unweighted analysis (B)

	C	G	H	W
Cattle related		0.269	0.014*	0.001***
Goat related	0.269		0.001***	0.001***
Human related	0.014*	0.001***		0.001***
Water related	0.001***	0.001***	0.001***	

The correlation between geographical area and variation of SGM abundence has been well proven in soil and water samples but there are no new evidence on animal product and shedding samples (Khera 2012, Pontiroli, Khera et al. 2013). No significant relationship was revealed using Weighted and Unweighted analysis in both PCoA plot and NMDS plot (Figure 5.16 & 5.17). The least similarity was between Middle and IDODI region compared to PAWAGA/middle and PAWAGA/IDODI regional pairwise analysis in both Weighted and Unweighted analysis (Table 5.7). This suggests that the geographical factor remains probably less important in explaining the composition of SGM and abundance.

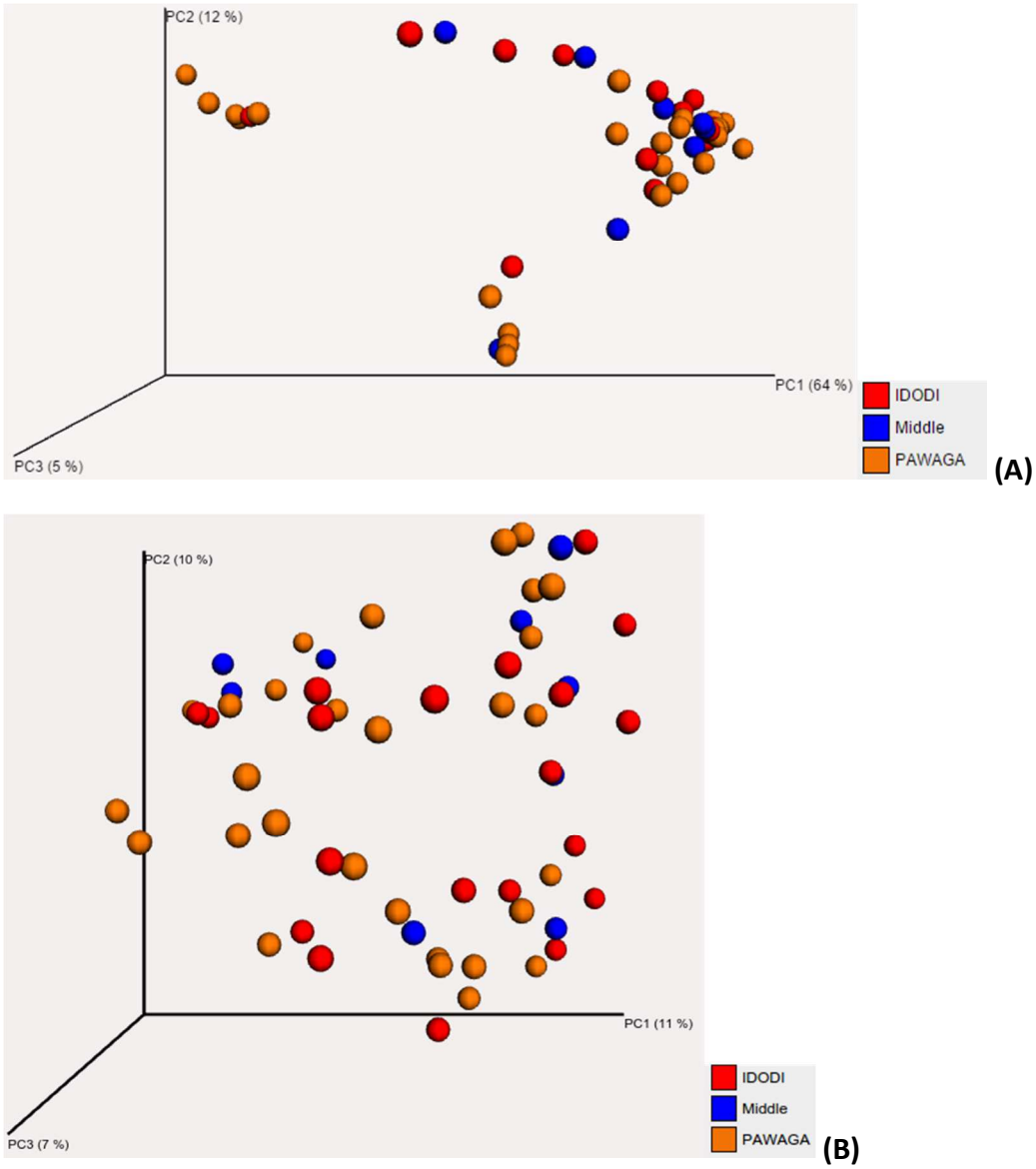
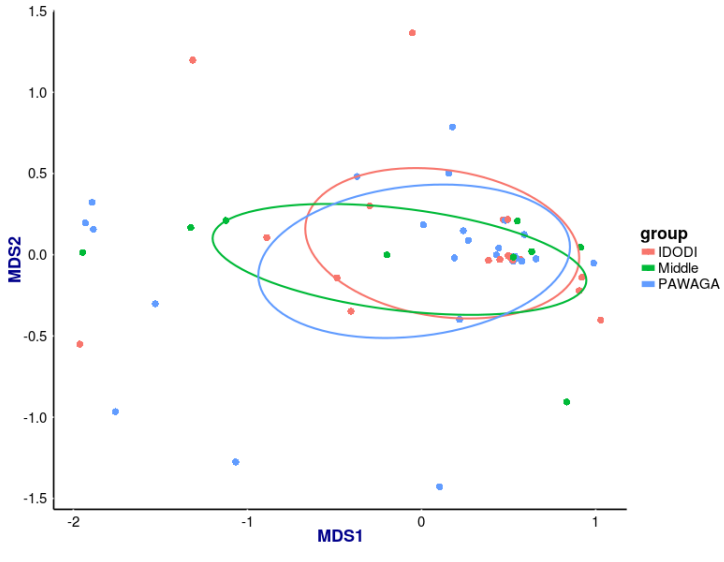
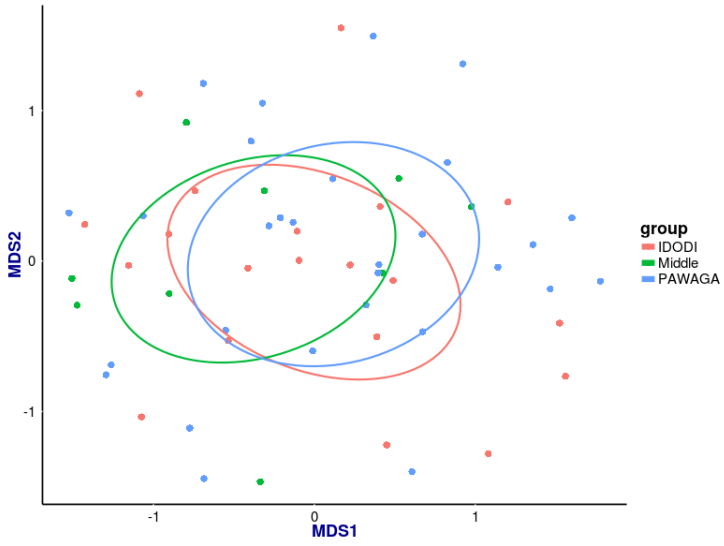


Figure 5.16 Comparison of different sampling regions in **(A)** Weighted analysis **(B)** Unweighted analysis.



(A)



(B)

Figure 5.17 Comparison of different sampling regions in **(A)** Weighted analysis **(B)** Unweighted analysis in NMDS plot.

Table 5.7 Comparison of different sampling regions. The significant difference is represented by P-value. * ANOSIM test (P-value < 0.05) in **(A)** Weighted analysis **(B)** Unweighted analysis.

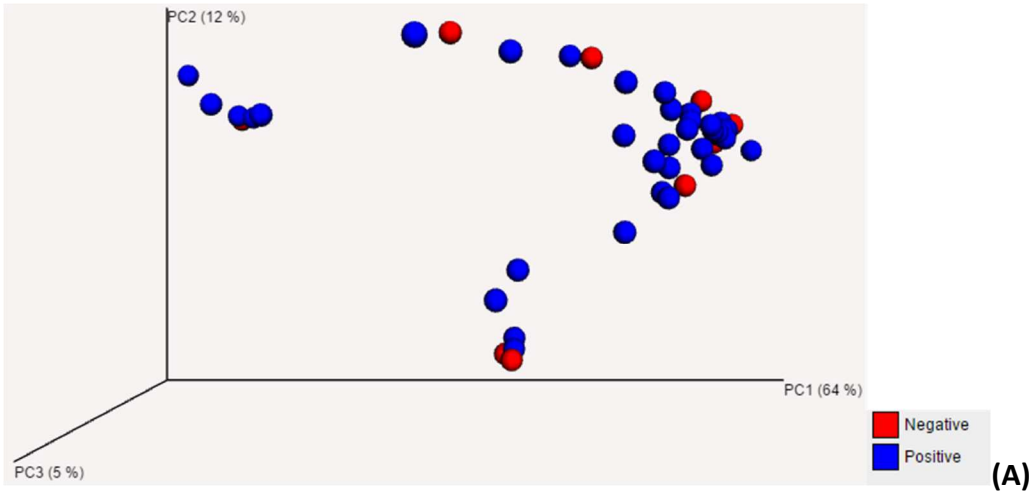
Pairwise comparisons for QIIME analysis (P-value) in Weighted analysis (A)

	IDODI	Middle	PAWAGA
IDODI		0.511	0.948
Middle	0.511		0.820
PAWAGA	0.948	0.820	

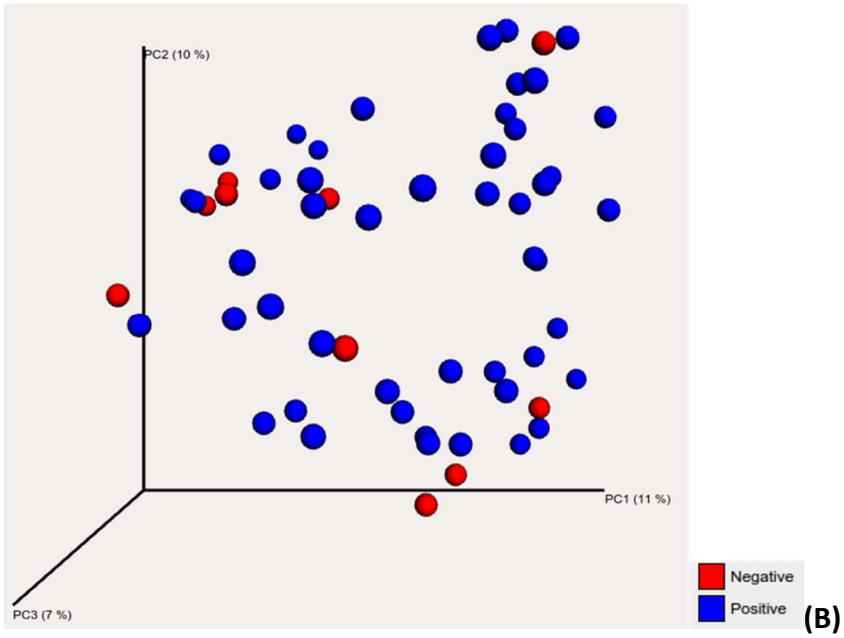
Pairwise comparisons for QIIME analysis (P-value) in Unweighted analysis (B)

	IDODI	Middle	PAWAGA
IDODI		0.037*	0.589
Middle	0.037*		0.145
PAWAGA	0.589	0.145	

No differentiation was described in the SGM pre-test classification (PCR and gel electrophoresis) using Weighted analysis in both PCoA plot and statistical analysis (Figure 5.18A& Table 5.8A). However the diversity of SGM abundance classified by SGM pre-test displays a slight segregation from positive to negative results using Weighted analysis in NMDS plot (Figure 5.19A). This indicates that the SGM composition in negative samples was only slightly different from positive samples. The statistical analysis in agreement shows that no significant difference between positive and negative SGM pre-test using both Weighted and Unweighted analysis. It suggests that the SGM primer was not sensitive in PCR detection and Gel electrophoresis to quantify the population of SGM and also not sensitive in determining in genus assignment or SGM variation level.

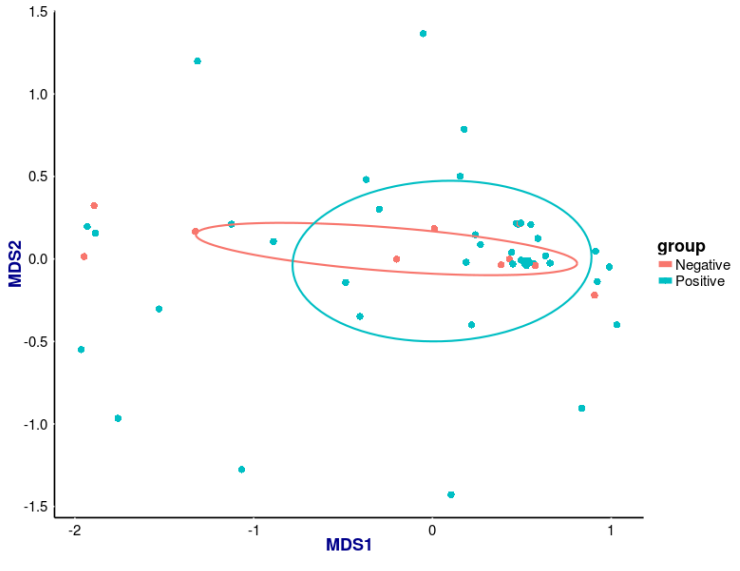


(A)

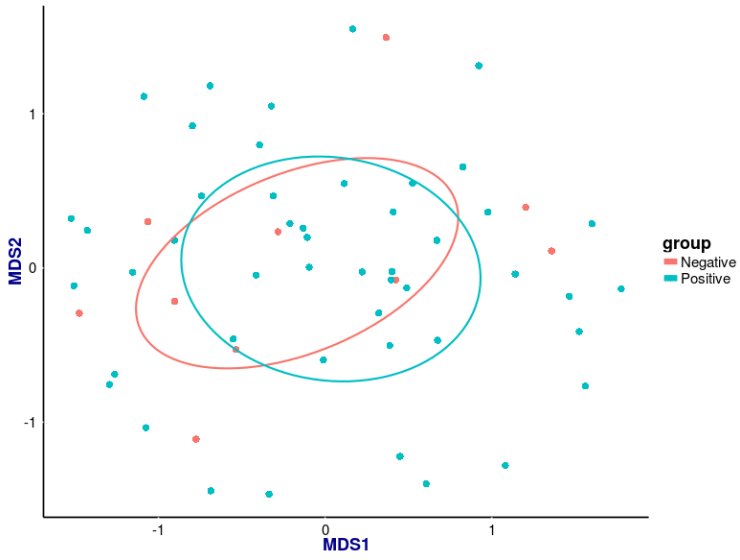


(B)

Figure 5.18 Comparison of two SGM pre-tests in **(A)** Weighted analysis **(B)** Unweighted analysis using PCoA plot.



(A)



(B)

Figure 5.19 Comparison of two SGM pre-tests in **(A)** Weighted analysis **(B)** Unweighted analysis in NMDS plot.

Table 5.8 Comparison of two different SGM pre-tests. The significant difference is represented by P-value. * ANOSIM test (P-value < 0.05), in **(A)** Weighted analysis **(B)** Unweighted analysis.

Pairwise comparisons for QIIME analysis (P-value) in Weighted analysis (A)

	SGMtest positive	SGMtest negative
SGMtest positive		0.304
SGMtest negative	0.304	

Pairwise comparisons for QIIME analysis (P-value) in Unweighted analysis (B)

	SGMtest positive	SGMtest negative
SGMtest positive		0.164
SGMtest negative	0.164	

5.4. Discussion

High throughput pyrosequencing was developed three decades ago (Ronaghi, Uhlen et al. 1998) but was not modified until 1999 to be utilised for microbial sequence identification and antibiotic resistance genotyping (Olive and Bean 1999, Boxer 2000, Clarke 2005). The MTBC were reported to be successfully distinguished from other bacterial cells using pyrosequencing with the highly specific RD primer (Novais, Borsuk et al. 2008). In addition MTBC was identified using same sequencing methods with highly conserved 16S rRNA primers from clinical samples (Heller, Jones et al. 2008).

A previous paper first reported the analysis of EM diversity using pyrosequencing with the same APTK primer and this diversity distinguished MTBC members from other SGM members distributed in water and soil samples (Pontiroli, Khera et al. 2013). This study examined the diversity of *Mycobacterium* communities with same primer using same 454 pyrosequencing but from different environmental samples such as faeces, household dust and sediment. The comprehensive pyrosequencing approach has enabled biogeographical analysis of diversity and composition of mycobacteria as determined by OTU and phylogenetic dissimilarities in QIIME analysis. The targeted pyrosequencing run specifically for SGM enabled a rare insight into the group community structure and diversity. It was of interest to assess whether the SGM group presented any separate trends in different sample sets or regions.

The QIIME sequencing analysis was applied after 454 pyrosequencing for diverse sample sets which includes sample types, organism related sample types, geographical region and SGM primer pre-test. The multiples QIIME analyses were combined with different diversity and statistical analyses provided a powerful tool to investigate the connection within samples and their environmental factors.

The specific 16S rRNA pyrosequencing primer for SGM bacterial group was sensitive for 454 pyrosequencing while amplification of SGM amplicon. The length of amplicon after amplifying was 420~430bp as expected. Three quarters of the dataset were assigned to the *Mycobacterium* genus when the sequence similarity threshold was set at 97%. However the occurrence of unassigned group was an issue for sequencing analysis. The relatively large unassigned groups may be due to random nucleotides insertions and deletions in a sequence read occurring and leading to mismatch during

taxonomic assignment. These errors were then not recognised and eliminated during denoising and chimera removal efficiently (Prabakaran, Streaker et al. 2011). Alternatively, as the assignment relied on databases that have mainly been produced from collections of isolated microbial organism reference, these databases are still not comprehensive and in particular lack sequences from whole bacterial communities, especially in the environment.

The introduced random sequencing errors were addressed via selection and elimination of smallest OTU with only one or two sequences to prevent these errors (Tedersoo, Nilsson et al. 2010). However there are still no efficient solution to second issue until whole environmental bacterial species sequence are identified and included in curated databases.

As a large proportion of sequence were unassigned, the alpha and beta diversity analysis would have been affected as these are still based on OTU classification, however, removing the potentially fake OTU improves the efficiency and accuracy of OTU classification on downstream diversity analysis (Hill, Walsh et al. 2003). Alpha diversity was examined by looking at OTU richness and abundance in each bacteria species or sample sets based on species accumulation curve and two asymptotic estimators. The assemblages of SGM in environmental samples are similar using alpha diversity analysis for variation in richness and abundance when classified according to sample type. Nevertheless the cattle related samples including cattle faeces and cattle boma soil are significantly different from rest of sample sets consisting of goat, human and water related samples. It reveals a fact that the cattle shedding samples contain bacterial communities that are unique and variable in

abundance, especially in SGM composition compared to other two organisms in shedding and environmental samples.

The spatial factor was probably a crucial component in the distribution of mycobacteria levels using alpha diversity analysis. These samples from the south region, IDODI division, are significantly less rich in mycobacterial abundance among the Iringa province. This mycobacterial difference indicates that the spatial factor plays an essential role in influencing bacterial composition in organism shedding samples even though the sample types are the same, only the region differs.

Beta diversity reflects that the effect of each variation, particularly environment or spatial factors on bacterial composition. It indicated how the SGM bacteria separate but normally it was subtle separation. Therefore the other distance matrix and multivariation analysis shall be introduced combined with statistical analysis to present the dispersal and distribution of SGM mycobacterial composition.

Complementary multivariate approaches facilitated the determination of the relative roles of environmental variation. Sampling faeces, soil and water environment in these pastoralists villages across a range of geographical regions provide complete analysis of the correlation and distribution of mycobacteria among human, livestock and environmental reservoirs. There was a high similarity of OTU shared between boma soil/household dust, faeces/boma soil and boma soil/sediment samples in OTU abundance level. It was noticed that there were significantly different bacterial communities among these three pairwise but shared similar abundances of mycobacteria. It provides a suggestion that strong interaction between environmental reservoirs and livestock was observed even though each sample type

contains their own unique and rare OTU but share the majority of their OTU at similar abundance. The abundance of OTU in water samples was not similar to other sample types and at the SGM level. Differences were observed among household dust/faeces and household/water in abundance analysis but not between dust/boma soil demonstrating the strong relationship among household dust and boma soil sample types was provided because these two sample types share same place in the village.

There were observed differences in human shedding/water related samples in OTU abundance and presence/absence tests. In addition there were differences in human shedding/goat shedding samples in both abundance tests and OTU presence/absence tests. This estimation of the significant difference occurred in human shedding samples compared to other samples at the level of SGM. Nevertheless water related samples are all associated with cattle in both OTU abundance and presence/absence test. Therefore the results suggest that SGM community structure and diversity present strong correlation between cattle shedding and water source rather than water and faeces/boma soil samples because faeces/boma soil still contained goat related shedding. Goat faecal and related boma soil samples showed high differentiation from water/sediment samples in OTU abundance and presence/absence level.

No differences were observed according to regional classification. This estimation displays that the region was not the main factor of SGM distribution within sample collected from different regions but it does slightly influence the SGM variation from the diverse regions.

Pyrosequencing data illustrated that there was no difference in statistical analysis among samples which were tested using PCR with SGM 16S rRNA primers and identified by gel electrophoresis as positives and negatives. It reveals the issue regarding to the specificity, sensitivity and accuracy of identification by gel electrophoresis. Unlike pyrosequencing, the band of target DNA products appeared in the gel electrophoresis is only obtained if the aliquot of PCR products are > 20ng of DNA products but cannot be separated from different bacterial species if the primer was universal or multiple species primer. Therefore it was not suitable for diversity analysis. Pyrosequencing allows diversity analysis by targeting universal or multiple bacterial species via OTU classification and identification. This next generation sequencing now can be used for large scale environmental screening, establishment of diversity analysis in SGM bacterial level and correlation among human living space, their livestock shedding and environmental reservoirs.

NMDS plots demonstrated that almost no segregation between sample types in both Weighted and Unweighted analysis. It suggests that the high proportion of unassigned species and insertion/deletion noises limited the differentiation of sample types and NMDS plot is ranking distance matrix which reduce the efficiency of variation between sample types. Therefore Oligotyping and MED shall be introduced to compare the efficiency of segregation, noise removal and species classification.

Chapter 6 The comparison between QIIME and novel sequencing Oligotyping for diversity analysis and environmental screening

6.1. Introduction

The QIIME sequencing analysis relied on a similarity threshold for clustering several similar sequences into OTU and classification via assignment of each representative DNA sequence to the reference sequence provided in the curated database (Caporaso, Kuczynski et al. 2010). This clustering based method reduced time and labour as only the reference sequences were used to compare to reference databases instead of every single sequence in the whole dataset. In addition these methods were used for diversity analysis including both alpha and beta diversities for building correlation among samples sets to environmental factors. However there are critical limitations of QIIME analysis: similarity threshold selection and resolution of reference database for species identification or diversity analysis.

The OTU identification strategy however, has critical limitation due to an incomplete taxonomic database to resolve diversity description, especially for the samples collected from high-diversity environments (Eren, Maignien et al. 2013). Another limitation was the relative similarity threshold provided to minimise the effect of random sequence errors on sequence analysis but a lack of references to identify whether they were real sequence errors or subtle differences in 16S rRNA (Kunin, Engelbrektson et al. 2010). In addition the random errors increase the likelihood of creation of fake OTU because a unique sequence causes these fake groups to result in bias in sequence identification and diversity analysis.

The efficiency of taxonomic identification of QIIME analysis is still limited to the sequence reference numbers contained in a recently curated reference database comprised of uclust (QIIME default), NCBI BLAST and RDP. The effect of taxonomic assignment subsequently influenced the downstream analysis. Another issue raised was the representative sequences selected from each OTU. The user defined selection of representative sequences was variable and the default selection in QIIME was to filter the most abundance sequence in each OTU to represent this OTU. Here the less abundant sequences were neglected as noise or forced to be classified to the same taxonomic unit as the representative sequence. The best reference database was still controversial and debated but the BLAST classifier was reported to provide taxonomic classification results with high confidence estimates compared to other two databases, especially for full length to less than 400 bp of 16S rRNA segment (Wang, Garrity et al. 2007, Taib, Mangot et al. 2013).

Segregation of SGM species faced a bottleneck due to the high similarity in whole genome and highly conserved region 16S rRNA region among each members of SGM species (GP 1984). The MTBC members are > 96 % similar in their whole genome and identical in the 16S rRNA. In addition the MAC members are consisted of 99.7 % similarity in the 16S rRNA region (Chaves, Sandoval et al. 2010). This raises problems if the similarity threshold was defined lower than 99.7 % and all MAC members were clustered into one OTU and not differentiated using QIIME for diversity analysis.

Oligotyping is another robust sequencing analysis method based on single sequence position variation compared to whole sequence similarity cluster (Eren, Maignien et al. 2013). This computational method was reported for application of environmental

samples sequencing analysis such as sewage water samples (McLellan, Newton et al. 2013) and pathogens distribution in the gut microbiome ecosystem to reflect human preference (Eren, Sogin et al. 2015). This algorithm based on entropy analysis filters the occurrence of the sequence errors effectively via accumulation of sequence position variation (Equation 2.4). The high abundance of sequence variation was able to be selected and retained but discriminated between those and the occurrence of random sequence which is identified as noise in the entropy analysis. Unlike the similarity threshold and classification based approach, this principle improved the massive dataset quality control and noise filtered during diversity analysis without discrimination of subtle sequence variation.

Through entropy analysis and Oligotyping, the resolution of species identification reached to only 0.2 % variation between different members in a highly conserved 16S rRNA region (Eren, Zozaya et al. 2011). This efficiency of species identification was suitable for segregation of SGM bacteria for environmental screening, especially for MAC members without the use of side experiments such as qPCR with specific primers to identify or quantify each MAC member.

Beta diversity was able to be estimated while exploring environmental samples with oligotyping. The species distribution obtained from oligotyping was plotted using NMDS for visualisation of correlation via similarity or dissimilarity of each sample within dataset. The NMDS in Oligotyping and MED, in comparison to QIIME in chapter 5, was more flexible for representing the relationship between objects. Unlike PCoA plots as linear correlation, NMDS plot were presented with rank order correlation following Bray-Curtis dissimilarity index for measure species composition between

metagenomes (Magurran. 2004). The rank ordination was more appropriate for reducing the effect of meta functional/taxonomic for comparative analysis. Statistical analysis ANOSIM was used to compare the similarity between different sample types.

Oligotyping relying on entropy analysis was limited in applicability to environmental screening datasets by the two components:

1. Oligotyping required the high accumulation of entropy in sequences for species identification between two closely related taxa species.
2. Oligotyping is still based on taxonomic classification in the final step for species identification.

In the recent study, the relationship of diversity analysis in each sample still relied on distribution of bacterial species in samples. However the result was not completed until each single bacterial species were identified in the sample sets. The problem was raised regarding species identification because this identity system still relied on the comparison unknown species genome from environmental samples to curated reference database but with poor resolution of reference genome to identify most of bacterial species from environmental screening. The novel remedy developed to resolve these issues named as Minimum Entropy Decomposition (MED) aimed to skip the step of species classification using a database and this approach only focused on the variation in amplicons. The amplicons were selected by decomposing each high entropy sequence location to each group with only one entropy sequence location and clustered as the Node group (Figure 1.5). The differentiation or similarity of each samples depended on the number of Nodes instead of species identification and oligotypes. This novel approach was sensitive for describing the diversity of closely

related organisms in high throughput sequence analysis, especially in environmental screening while avoiding the use of species reference databases (Eren, Morrison et al. 2015).

6.2. Aims

1. To investigate the bacterial abundance and distribution in environmental reservoirs via 454 pyrosequencing and Oligotyping analysis.
2. To compare the Qiime, Oligotyping and MED approach for diversity analysis in environmental screening.
3. To correlate human, livestock and environmental communities via environmental sample screening.

6.3. Results

6.3.1. Entropy analysis and identification of SGM species abundance distribution

The Shannon entropy applied in the environmental screening identified 14 interesting sequence positions that were selected for sequence classification and identification (Figure 6.1). The three most information rich nucleotide peaks distribute in the 269th, 58th, and 101st position in this alignment. These positions represented the highest variation among these sequences to form diverse oligotypes using oligotyping analysis. In total 320 oligotypes were generated using these positions of interest and a diverse proportion of oligotypes were distributed in each sample (Figure 6.2). However a proportion of these oligotypes were at low abundance and added noise to the analysis. The elimination of oligotypes based on

abundance can therefore simplify the information and focus on the majority of oligotypes in each sample. Distribution of minimum percent abundance of an oligotype > 20 % in total was selected in each sample (Figure 6.3). It was notable that the distribution of oligotypes within sample types was totally different, especially in faecal, water and other samples. The 320 oligotypes were reduced to only 11 major oligotypes in this alignment. The elimination aims to reduce time consumption and obtain the majority of bacterial species distribution but lose subtle information without affecting the analysis. The time period, for identification, was shortened from 20 hour for identification of 486 oligotypes to only 10 minutes for 11 oligotypes.

The abundance selection results in subtle information loses (Figure 6.4 & 6.5) and sample differentiation in diversity analysis (Figure 6.6). The distribution of *Mycobacterium* in this alignment within both selections shows that there is a different proportion of *Mycobacterium* genera and species in all abundance selections compared to > 20 % abundance selection, especially at species level (Figure 6.4 & 6.5). The absence of MTBC members was observed in > 20 % abundance selection but they appeared in the all abundance selections because the proportion of MTBC was identified at < 20 % and eliminated by algorithm. In addition the proportion of two major groups, *Mycobacterium asiaticum* and *Mycobacterium intracellular* are both slightly reduced from > 20 % abundance to all abundance selection. It was suggested that the only minorities of species are influenced by abundance selection but not the prevalence of majorities.

The effect of diversity analysis was observed in Figure 6.6 within two selection, particularly in sample differentiation. Household dust and sediment samples

segregate from boma soil samples in the presence of all species compared to only majority group selection. Nevertheless there was no segregation in faecal samples in both selections because the mycobacteria species present in faeces are oligotype diverse from other samples (Figure 6.2 and 6.3).

The issue found in QIIME was the group of uncultured bacteria genera and species because of the resolution of bacterial reference species in the curated database. Oligotyping reduced the proportion of uncultured bacteria from 13 % in QIIME to only 3 % but this ambiguous cluster still persisted in this alignment (Figure 6.4 & 6.5).

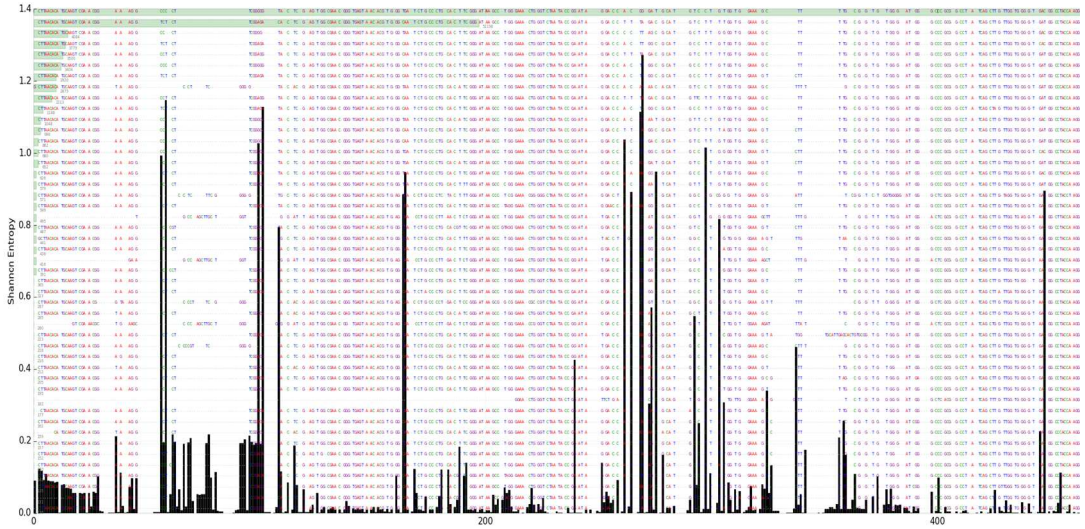


Figure 6.1 The entropy analysis plot indicates that the sequence location of high entropy in the dataset. The information-rich position located in 56th, 58th, 99th, 101st, 108th, 163rd, 164th, 261st, 264th, 268th, 269th, 275th, 297th and 303rd were selected as represent position for classification of oligotypes.

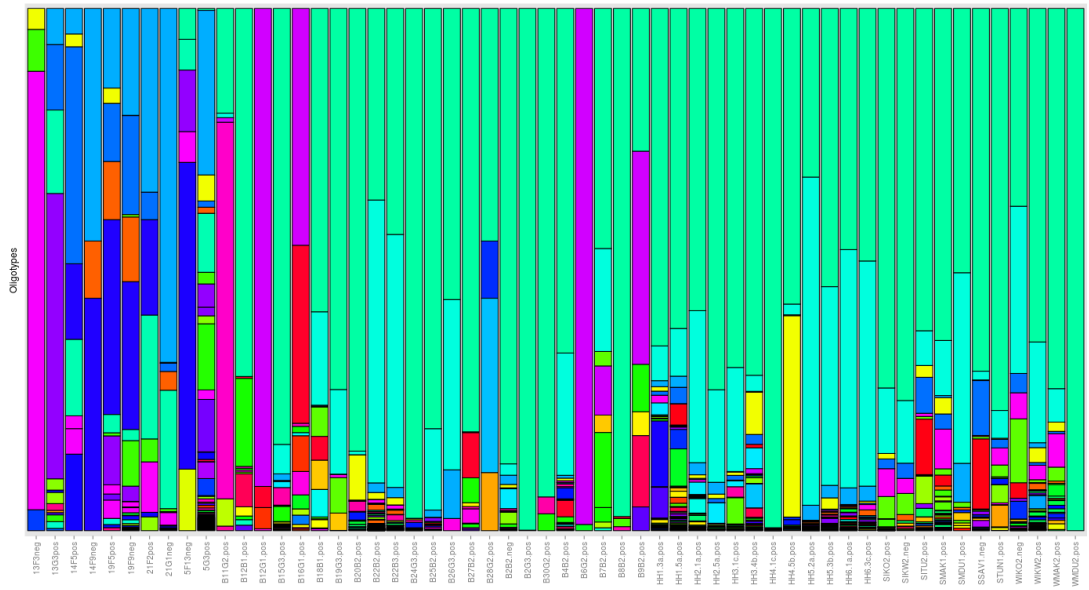


Figure 6.2 Distribution of 486 oligotypes in each samples. All oligotypes in each sample were plotted in this Figure. Different oligotypes were highlighted in colours.

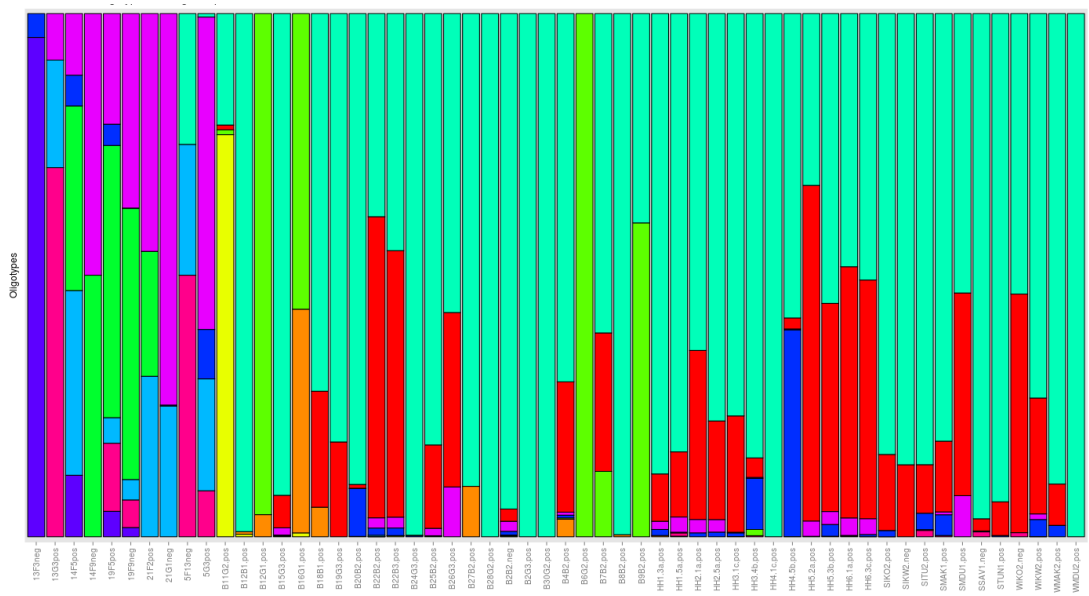


Figure 6.3 Distribution of 486 oligotypes in each samples but the abundance of oligotypes > 20 % in this alignment were eliminated in this Figure. Different oligotypes were highlighted in colours.

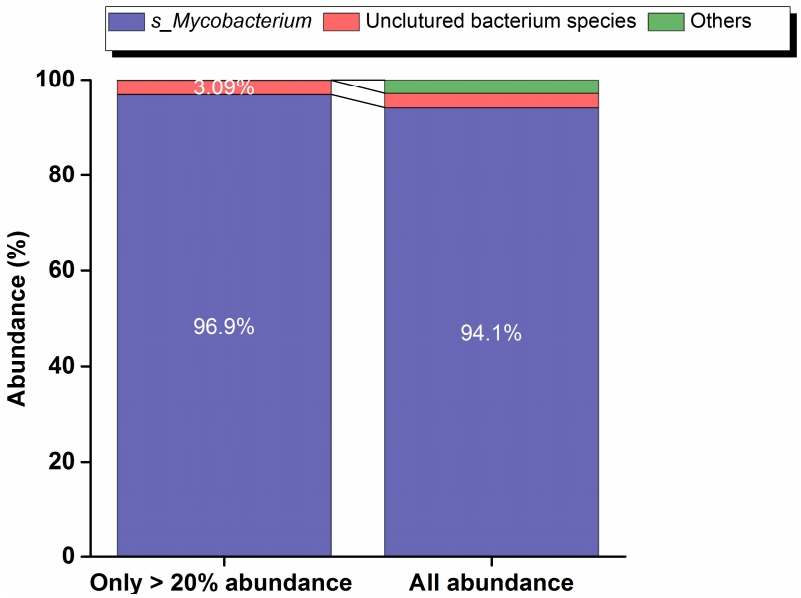


Figure 6.4 The distribution of Mycobacterium at species level within both abundance selections.

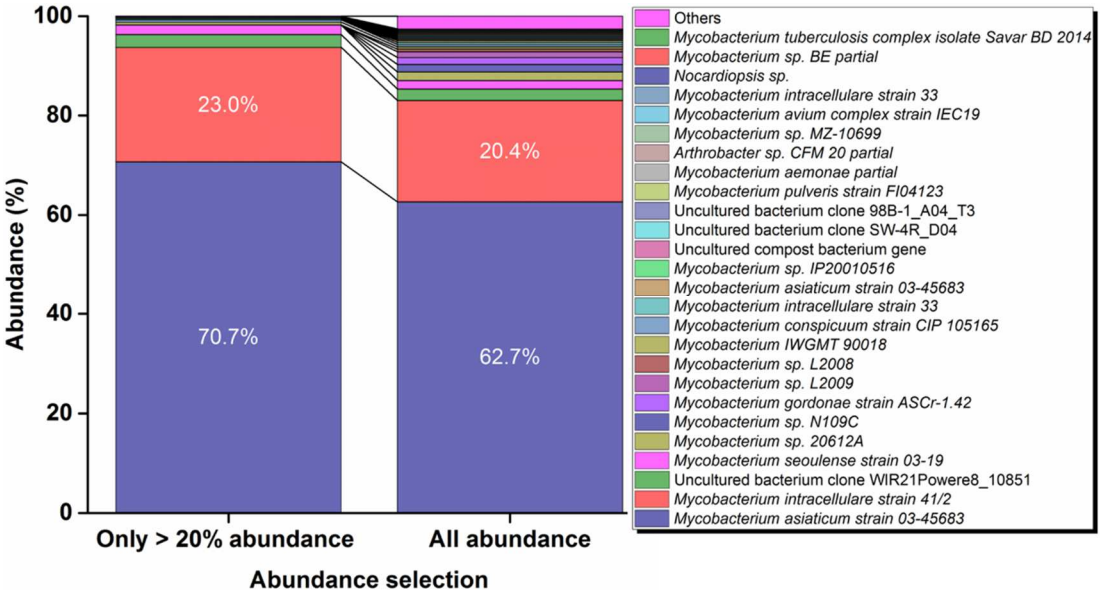


Figure 6.5 The distribution of Mycobacterium in species level within both abundance selection.

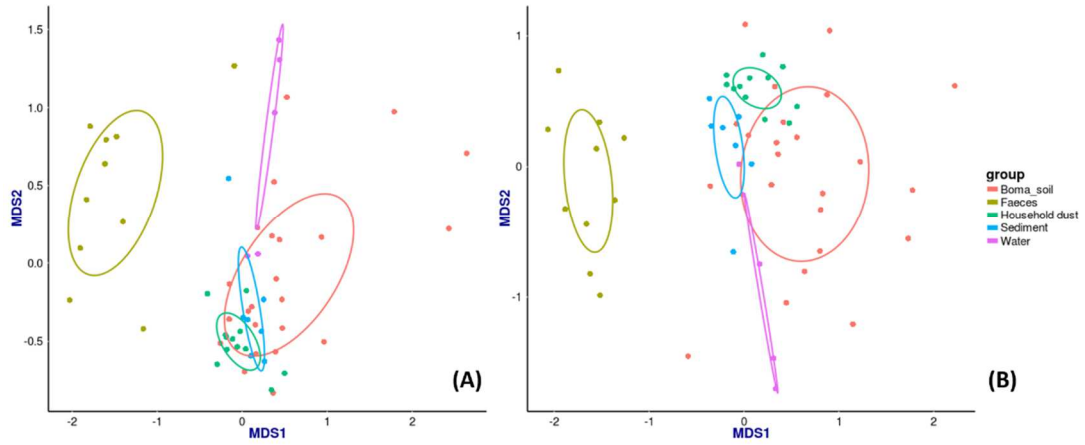


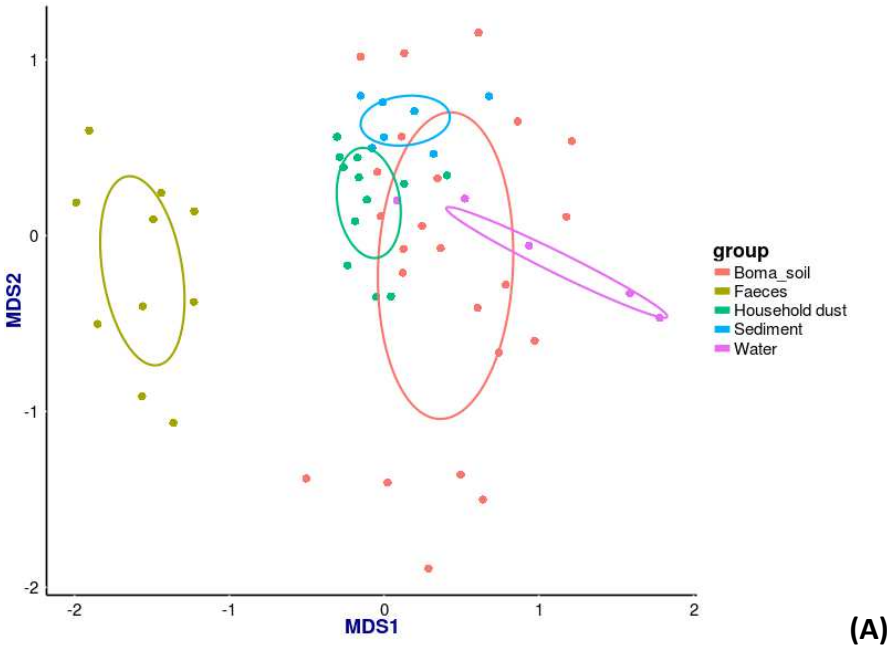
Figure 6.6 The NMDS plots show the comparison between the abundance of oligotypes > 20 % in total (A) and all abundance selection (B).

6.3.2. Diversity analysis using Oligotyping and MED

The distance ranking NMDs plots and statistical analysis compared the diversity between sequence variable MED approach and species classification based Oligotyping analysis to understand the correlation of environmental samples sets and spatial distance. The distance ranking NMDs plot was based on the matrix percentage which was generated from Oligotyping and MED algorithms to indicate the abundance of each oligotype or Nodes depending on algorithms. The overlapping region in NMDs plots reveal the proportion of similarity of groups shared in diversity analysis among samples where the corresponding distance between each cluster represents the rate of dissimilarity among samples in NMDS plots.

The environmental samples clustered together except for faecal samples which separated in Oligotyping analysis (Figure 6.7A) and in MED approach (Figure 6.7B). This result indicates that the MED approach was able to differentiate each samples without species classification.

Significant differences between boma soil and other samples are detected using pairwise ANOSIM analysis in both Oligotyping and MED (Table 6.1A & B). The sequence distribution in water samples to sediment was extremely different (P-value = 0.001) in both Oligotyping and MED. The boma soil samples and household dust obtained in MED were not distinguishable (P-value = 0.067) but significantly different in Oligotyping (P-value = 0.001). In contrast to boma soil/household dust, the results in boma soil/water group presented the high difference in MED (P-value = 0.001) but similarity in Oligotyping (P-value = 0.184). Analysis suggests that the strong correlation was detected for boma soil and household dust sample in MED but not in species classification level because some species are still not recognised by known reference species databases resulting in bias in diversity analysis.



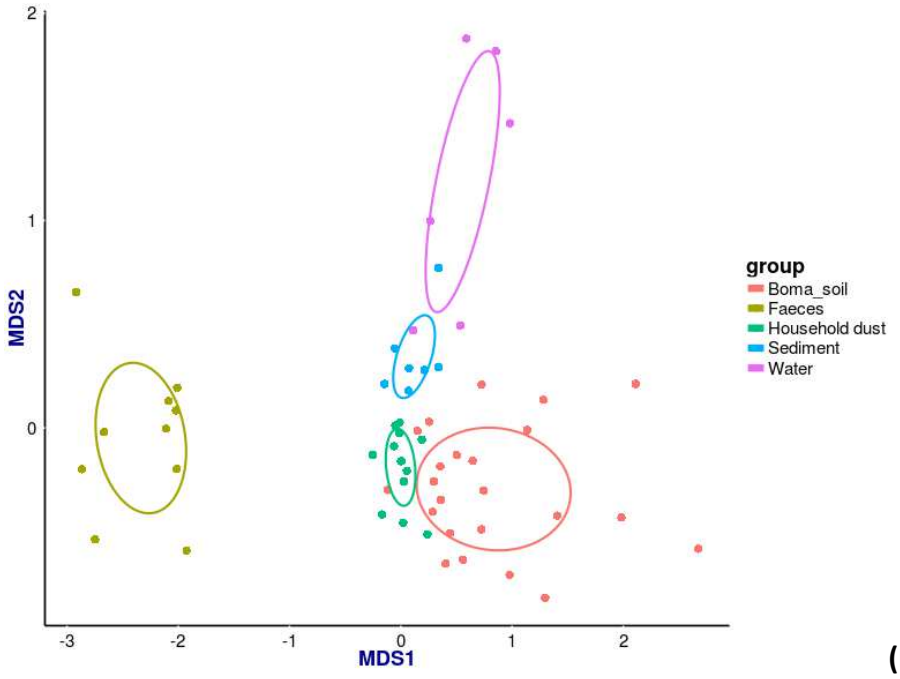


Figure 6.7 Comparison of different samples sets using **(A)** Oligotyping **(B)** MED analysis in NMDS plot.

Table 6.1 Pairwise comparison of different sample types. The significant difference is represented by P-value. ** ANOSIM statistical analysis (P-value < 0.01), *** (P-value < 0.001) in **(A)** Oligotyping **(B)** MED analysis.

Pairwise comparisons for Oligotyping analysis (P-value) (A)

	BS	F	HH	S	W
BS		0.001***	0.001***	0.001***	0.184
F	0.001***		0.001***	0.001***	0.001***
HH	0.001***	0.001***		0.001***	0.001***
S	0.001***	0.001***	0.001***		0.005**
W	0.184	0.001***	0.001***	0.005**	

Pairwise comparisons for MED analysis (P-value) (B)

	BS	F	HH	S	W
BS		0.001***	0.067	0.054	0.001***
F	0.001***		0.001***	0.001***	0.001***
HH	0.067	0.001***		0.001***	0.001***
S	0.054	0.001***	0.001***		0.005**
W	0.001***	0.001***	0.001***	0.005**	

All animal related samples overlapped, especially human related samples with goat related samples but were separate from water related samples in Oligotyping (Figure 6.8A). The MED in contrast shows only a quarter of water related samples overlapped other sample types and those that did overlapped human related samples (Figure 6.8B). A stronger relationship between water and human related samples was identified in MED than in Oligotyping. However more than half of cattle and goat related samples overlapped in both Oligotyping and MED. Although cattle and goat related samples share similarity in species classification level in both Oligotyping and

MED, complete segregation is observed between human and water related samples only in Oligotyping (Figure 6.8).

The strong correlation between goat and cattle related samples seen in ANOSIM analysis at Oligotyping (P-value = 0.891) and of MED (P-value = 0.800) is seen in the NMDS plot shown (Figure 6.8 & Table 6.2). It was noticed that there was a significant difference in goat compare to human and water related samples in both Oligotyping (P-value = 0.001) and in MED (P-value = 0.001). In addition there were no highly different in cattle compared to water related samples in Oligotyping (P-value = 0.052) in contrast to MED (P-value = 0.002). The result implied that the faeces and boma soil related to cattle share more similarity in the species classification level than at the sequence level to water/sediment. Both species classification and sequence level in goat related samples were distinguishable from human and water related samples but highly similar to cattle related samples. However MED showed water related sample overlapping human related samples although no similarity was presented in ANOSIM.

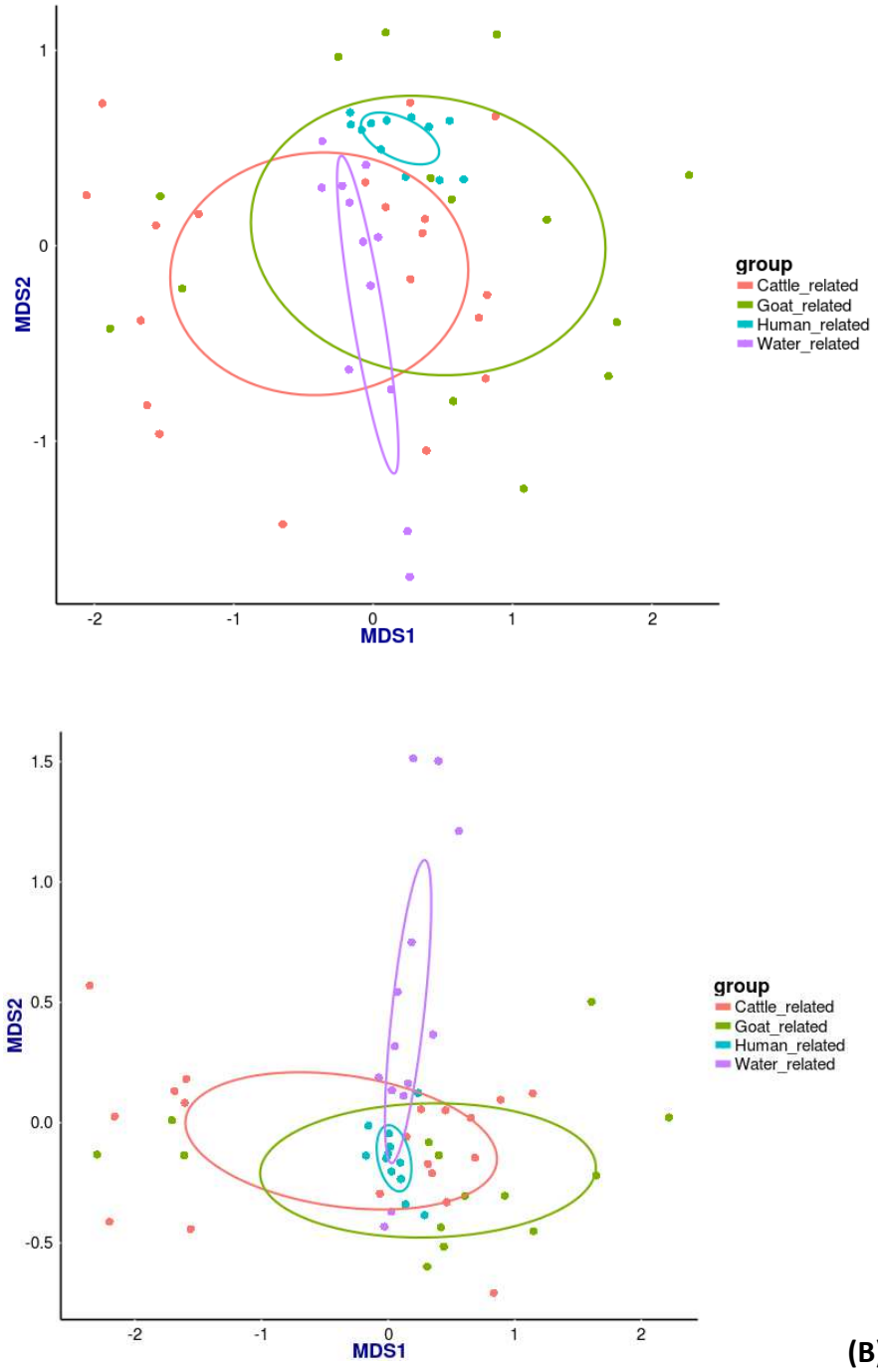


Figure 6.8 Comparison of different sample types using **(A)** Oligotyping **(B)** MED analysis in NMDS plot.

Table 6.2 Comparison of pairwise of different sample types. The significant difference is represented by P-value. * ANOSIM statistical analysis (P-value < 0.05), ** (P-value < 0.01) and *** (P-value < 0.001) in **(A)** Oligotyping **(B)** MED analysis.

Pairwise comparisons for Oligotyping analysis (P-value) (A)

	C	G	H	W
Cattle related		0.891	0.002**	0.052
Goat related	0.891		0.001***	0.004**
Human related	0.002**	0.001***		0.001***
Water related	0.052	0.003**	0.001***	

Pairwise comparisons for MED analysis (P-value) (B)

	C	G	H	W
Cattle related		0.800	0.012*	0.002**
Goat related	0.800		0.001***	0.001***
Human related	0.012*	0.001***		0.001***
Water related	0.002**	0.001***	0.001***	

The samples did not segregate by region in either Oligotyping or MED. High similarity within samples from these three regions occurred, especially in Middle and PAWAGA regions (Figure 6.9). The statistical analysis identified that the distribution of samples in species classification and sequence level in the IDODI region are highly differentiated from samples in both PAWAGA and Middle region in both Oligotyping and in MED than PAWAGA to Middle. This result suggests that the geographical factor in this study is probably not the main influence in SGM distribution.

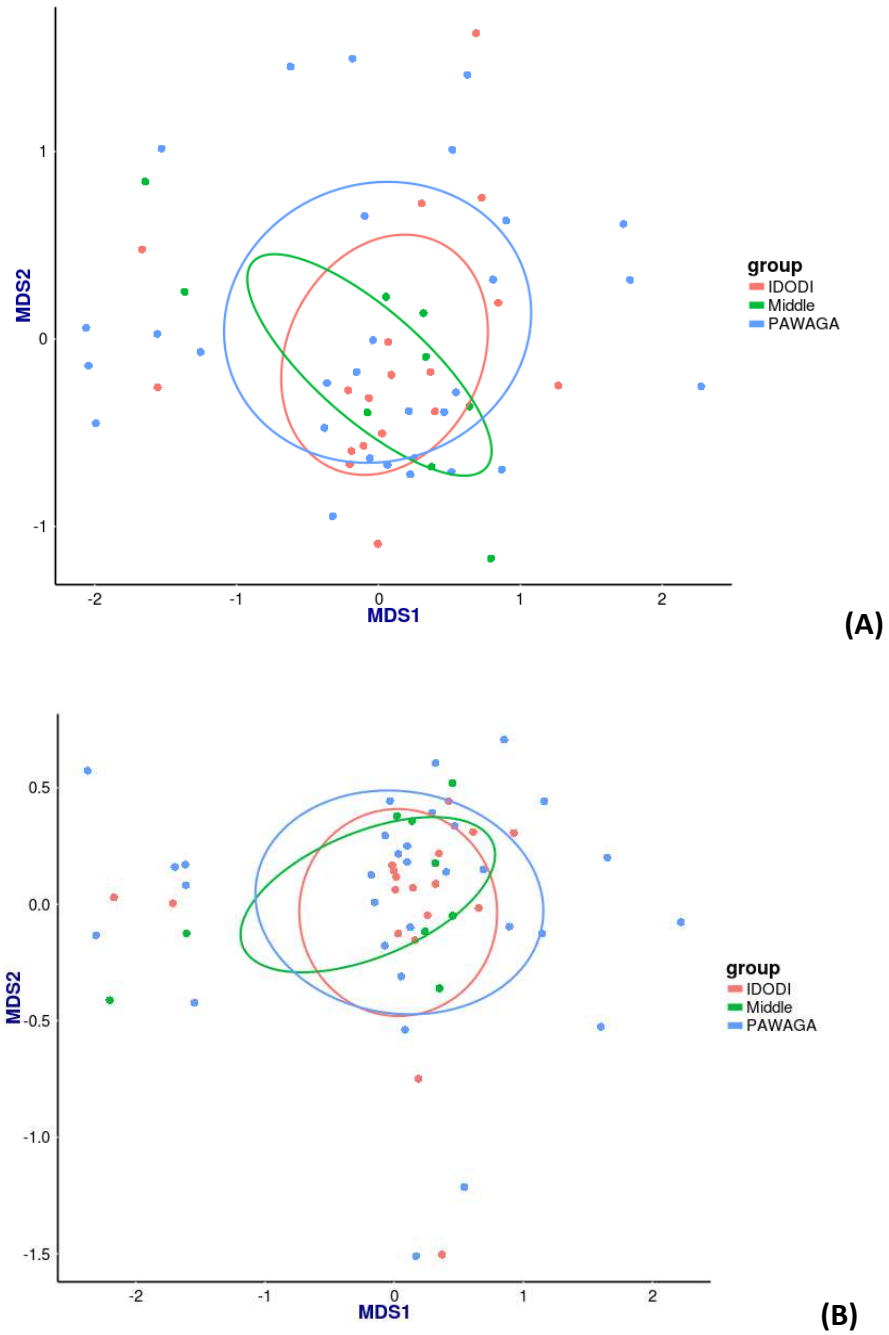


Figure 6.9 Comparison of different sampling regions using **(A)** Oligotyping **(B)** MED analysis in NMDS plot.

Table 6.3 Comparison of pairwise of different sampling regions. The significant difference is represented by P-value. ANOSIM statistical analysis in **(A)** Oligotyping **(B)** MED analysis.

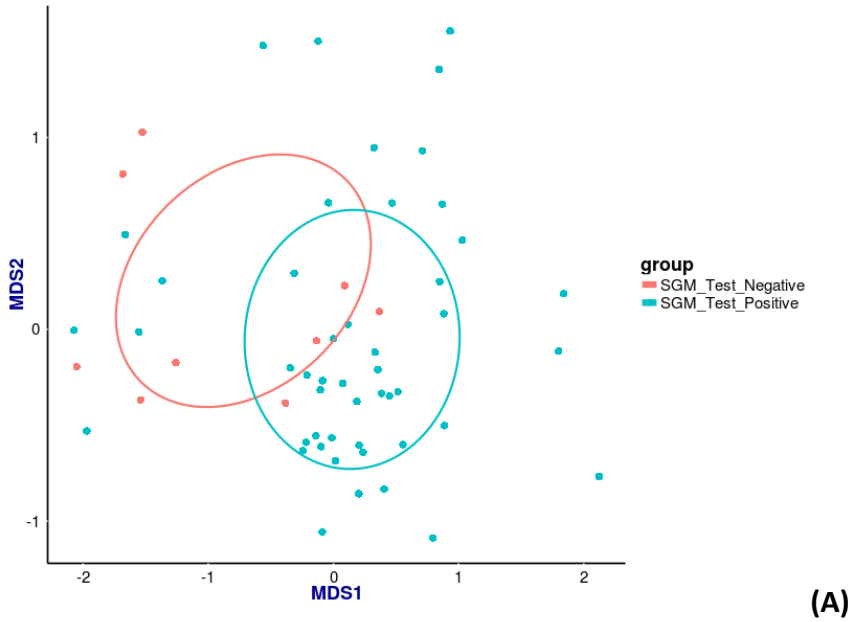
Pairwise comparisons for Oligotyping analysis (P-value) (A)

	IDODI	Middle	PAWAGA
IDODI		0.095	0.072
Middle	0.095		0.378
PAWAGA	0.072	0.378	

Pairwise comparisons for MED analysis (P-value) (B)

	IDODI	Middle	PAWAGA
IDODI		0.125	0.149
Middle	0.125		0.515
PAWAGA	0.149	0.515	

Less than half of the SGM test positive and test negative samples overlapped in NMDS plot in both Oligotyping (Figure 6.10A) and MED (Figure 6.10B). These results implied that the total difference in both samples sets can be detected in sample species but not in the distance ranking test. The ANOSIM analysis demonstrates that there are extreme differences between two SGM pre-test samples in both Oligotyping (P-value = 0.018) and MED (P-value = 0.019). These results imply that the positive and negative samples identified based on gel electrophoresis do have differences that can be detected at the species and sequence level.



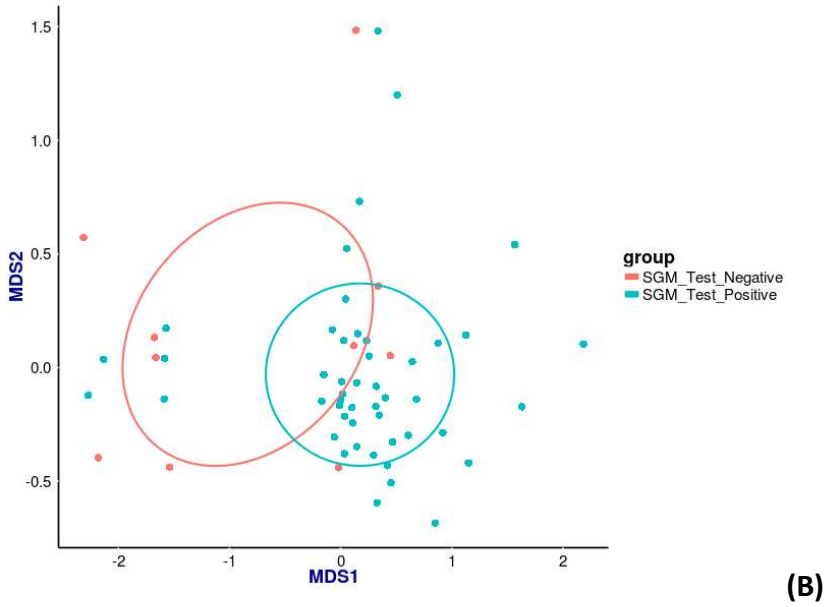


Figure 6.10 Comparison of SGM test sets using **(A)** Oligotyping **(B)** MED analysis in NMDS plot.

Table 6.4 Comparison of pairwise of SGM test sets. The significant difference is represented by P-value. * ANOSIM statistical analysis (P-value < 0.05) in **(A)** Oligotyping **(B)** MED analysis.

Pairwise comparisons for Oligotyping analysis (P-value) (A)

	SGMtest positive	SGMtest negative
SGMtest positive		0.018*
SGMtest negative	0.018*	

Pairwise comparisons for MED analysis (P-value) (B)

	SGMtest positive	SGMtest negative
SGMtest positive		0.019*
SGMtest negative	0.019*	

6.4. Discussion

This study has identified that Oligotyping was capable of elimination of a proportion of the unassigned sequences compared to sequence similarity based OTU classification algorithms. The species classification in QIIME was only identified to the genus level but not into species level.

The same curated database, BLAST, was applied in both QIIME and Oligotyping analysis for comparison but significant differences was observed in species classification and proportion of species identification. For example, 72.4 % of sequences were assigned to the *Mycobacterium* genus using QIIME whereas 94.1 % were assigned to *Mycobacterium* species in Oligotyping. The proportion of unrecognised species was reduced from 12.9 % in QIIME compared to 2.98 % in

Oligotyping. In addition the majority of SGM species in this alignment were identified as *Mycobacterium asiaticum* and *Mycobacterium intracellulare*.

These two SGM members have been identified as NTM human and animal pathogens. *M. asiaticum* was identified from monkeys as NTM in 1950 (Masson and Prissick 1956) and the first report of this pathogen being isolated from humans was in 1982 in Queensland, Australia (Blacklock, Dawson et al. 1983).

Mycobacterium intracellulare is a member of MAC belonged to SGM species. *M. intracellulare* as a pathogen is normally associated with immune compromised disease, in particular with AIDS (Karakousis, Moore et al. 2004).

The MTBC are unable to be segregated using 16S rRNA primers because of their identical 16S rRNA sequence region and their extremely low proportion in this alignment at only 0.1 %. The species classification reveals that the majority of *Mycobacterium* species from different environmental and livestock reservoirs were not MTBC. It is worth noting that the wildlife disease related mycobacteria species were detected from these environmental samples and persisting in water and sediments in the dry season even though there was a lack of MTBC. Opportunistic mycobacteria species can be detected in cattle and goat faeces and in human dwelling dust. It implies that the bacterial communities can be shared between water or sediment, livestock faeces and even the dust surrounding human houses.

The species classification was still uncertain because unrecognised bacterial species were still present in the Oligotyping analysis, even though they were at an extremely

low proportion. If it was possible to identify these unknown species were this may improve the accuracy of further diversity analysis.

The MED approach in contrast to Oligotyping provides efficiency by skipping the species classification matching with reference database. MED eradicates the proportion of unassigned sequences by choosing high entropic positions as a “Nodes” instead of classifying using OTU or oligotypes. The similarity between sample types was compared using the composition of different numbers of Nodes contained in each sample. MED cannot be used to identify bacterial species, however, MED can identify differences based on only a few nucleotide variations to differentiate samples. For example, MED showed strong differentiation between two deep-sea sponge cryptic species, *Hexadella dedritifera* and *H. cf. dedritifera*, which had been classified into the same OTU using QIIME analysis (Eren, Morrison et al. 2015).

PCoA plot based on distance matrix demonstrated the similarity or dissimilarity between samples types and more functional or taxonomic characters compared to NMDS plot based on ranking distance matrix. The statistical analysis ANOSIM test based on Bray-Curtis similarity matrix so the results are similar to the NMDS presented. Unlike the PCoA, NMDS clarifies relative position between samples types. For example, faecal samples are differentiated from other four samples in statistical analysis as in NMDS plot using the same method MED. This result suggests that ANOSIM statistical analysis can efficiently explain taxonomic data from the ranking distance based NMDS between samples types.

It was worth noting that a strong relationship was detected between cattle and water related samples only using Oligotyping, while cattle samples overlap goat, human and

water related samples in MED. A similar observation was described in chapter 3 where a high prevalence of Mb cells were present in cattle faeces, boma soil and water samples near households.

This study is the first to delineate Mycobacteria in diverse environmental samples sets via MED and use the differentiation to understand the relationship among samples by NMDS plots and statistical analysis. This sequence variation based analysis approach will be further employed to identify correlations between human disease and environmental reservoirs around their settlement.

Chapter 7 Shedding of *Mycobacterium tuberculosis* in Tanzanian household: the environment as a risk factor for infection

7.1. Introduction

TB is a widespread and devastating human pulmonary disease caused by Mtb, a pathogen which belongs to the slow-growing subgroup of the *Mycobacterium* genus (Kubica 1984). Mtb is a member of the MTBC which contains eight species including the animal pathogen, Mb (Etchechoury, Valencia et al. 2010). Infections with Mtb have resulted in a health burden worldwide and led to over eight million new cases of TB each year and more than a million deaths annually (Eurosurveillance editorial 2013). The majority of TB incidents occur in the BRICS group (Brazil, Russian Federation, India (2.0 – 2.4million), China (0.9 – 1.1million) and south-east Africa. The latter consists of Tanzania (80,000) and South Africa (0.4 -0.6million), and in Africa alone 27 % of global TB cases were recorded in 21 countries during 2012 (Eurosurveillance editorial 2013). Furthermore, around 2 % of the human TB prevalence identified in the world are due to the zoonotic pathogen, Mb, which is the causal agent of bTB (Etchechoury, Valencia et al. 2010). Recent estimates indicate that in Africa human infections attributed to Mb can exceed 10 %; in particular Ethiopia (16.3% from human specimen culture), Zambia (17.9%) and Tanzania (16% with TB cases of lymphadenitis) (Kazwala, Daborn et al. 2001, Shitaye JE 2007, Malama, Johansen et al. 2014). The reservoirs of bTB disease occur predominantly in livestock and wildlife with endemic disease within the cattle herds of UK, USA, Spain and New Zealand being attributed to wildlife reservoirs (Nugent, Whitford et al. 2012, Munoz-Mendoza, Marreros et al. 2013). However, previous studies revealed that

environmental contamination, caused by faecal shedding associated with infection provided a potential and indirect route for transmission of bTB infection (Courtenay, Reilly et al. 2006, Sweeney, Courtenay et al. 2007, Hayley King 2015).

Although TB has been classically presented as a clinical pulmonary disease transmitted by aerosol between human populations, some studies have found that the transmission route may be through consumption and ingestion of dairy products from domestic livestock and unsterilised water from contaminated environmental reservoirs leading to infection of TB (Roug, Perez et al. 2014). Recent epidemiological studies even indicated TB and other NTM can no longer to be considered as purely a disease confined to humans and livestock, but able to spread with environmental factors such as soil and dust particles from barns (Eisenberg, Nielen et al. 2010).

Our approach has been to develop molecular diagnostic methods to enable rapid, accurate and quantitative analysis of environmental contamination by Mtb. We have validated a method for qPCR assay of Mb suitable for soil, faeces, and other environmental sample types (Pontiroli, Travis et al. 2011, Travis, Gaze et al. 2011).

The high similarity between Mb and Mtb and difficulty of distinguishing these from MTBC provides a challenge for molecular diagnostics (Garnier, Eiglmeier et al. 2003). However, whole-genome DNA microarrays established RD and provided an efficient and specific target region for segregation of each MTBC member (Halse, Escuyer et al. 2011). The standard assay developed for environmental monitoring relied on the use of an RD-based primer which hybridises to a part of the RD4 deletion region (RD4 scar) (Sweeney, Courtenay et al. 2007). This holistic approach was used to establish faecal shedding levels in a wildlife population in the UK (Travis, Gaze et al. 2011). In

the current study we present application of qPCR analysis of shedding both Mb and Mtb in environmental samples collected in Tanzania as part of a wider study on the epidemiology, risk factors, and transmission dynamics of Mtb and Mb between humans and animals in a well-defined rural population in Tanzania.

Dust samples were collected from households that were TB case positive in Tanzania during two discrete seasons; in September during the dry and during February in the wet season. 135 TB-case households and 114 control households in same area were identified by the NIH project (Hopewell 2011). Control households were not identified as having active TB cases. Household samples consisted of livestock faeces and dust from food preparation areas, bathroom facilities and livestock housing area. The data indicated that Mtb shedding was detected in household dust but no Mb was detected in either dust or animal manure. This study highlights the potential importance of environmental reservoirs in the transmission of human TB.

7.2. Aims

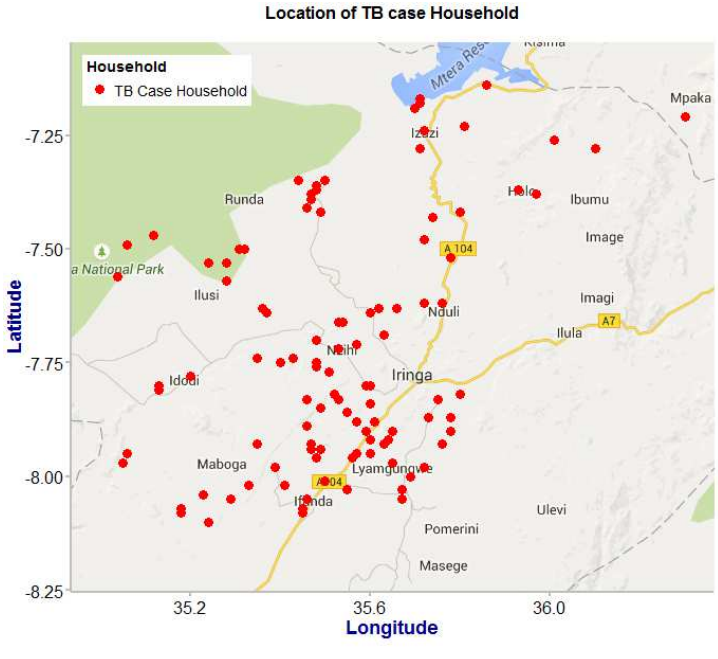
1. To quantify the number of Mtb contained in Tanzanian household dust samples and compare Mtb prevalence in the TB case household and control household dust samples
2. To introduce spatial analysis to investigate the distribution of Mtb in the study region.
3. To estimate the correlation between TB proportion and Mtb prevalence.

7.3. Results

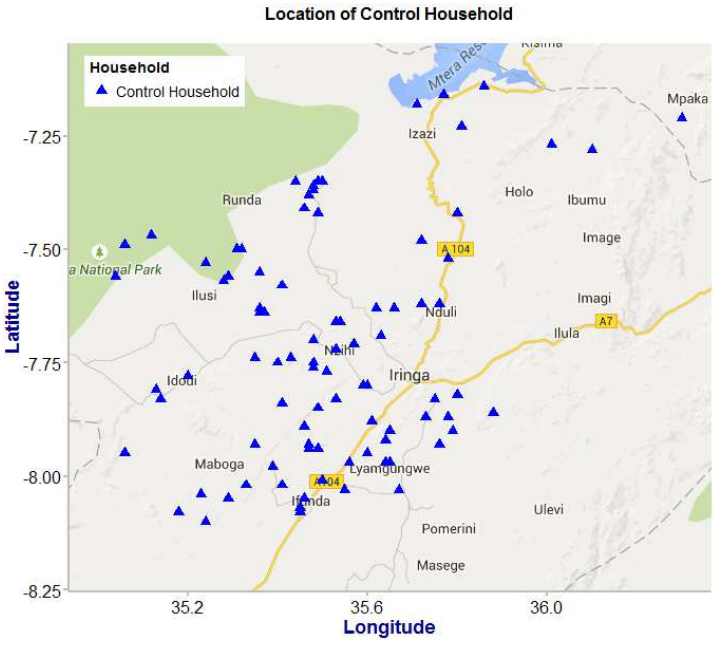
The spatial study showed the distribution of agriculturist households with and without TB cases located in Iringa region close to Ruaha national park and the local water supply, the Ruaha River. The 114 control households selected from households without TB patients, which were close to 135 TB case households enabled a comparison of Mtb distribution in the two sets of household dusts (Figure 7.1).

Mtb was detected in dust from 44.4 % TB case households compared to control households where 17.5 % were TB positive (Table 7.1).

The pronounced negative skew in the distribution of Mtb positives from TB case households was significantly higher than positive skewed distribution of control households (Figure 7.2). However there was no significant difference (P-value > 0.05) in terms of Mtb genomic equivalents in household dust between the two seasons (Figure 7.3).



(A)



(B)

Figure 7.1 TB case individual households (A) and Control households (B) in the study region of Tanzania

Table 7.1 The proportion of Mtb positives via dust from TB case and Control households

	TB case Household	Control Household
Number of dust samples	**135**	**114**
Positive	**44.4%**	**17.5%**
Negative	**55.6%**	**82.5%**

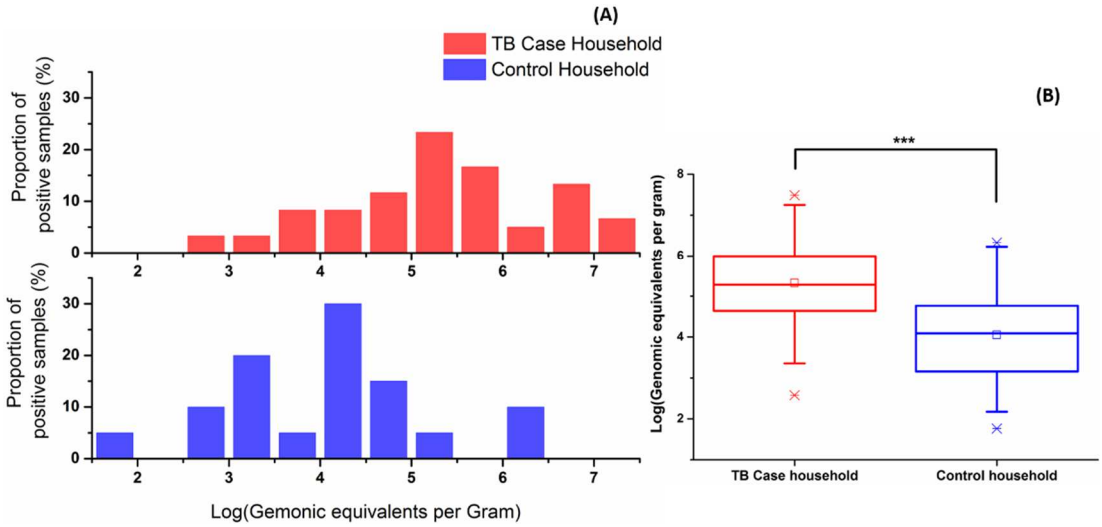


Figure 7.2 Mtb prevalence (A) and the range (B) in household dust over the two household sets. *** Mann Whitney test (P-value < 0.001).

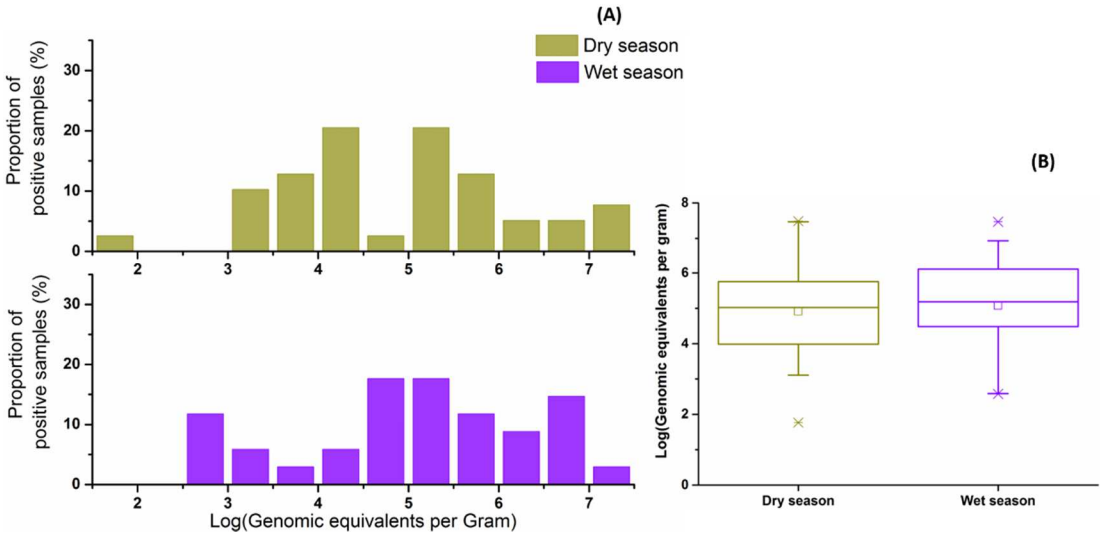


Figure 7.3 Mtb prevalence (A) and the range (B) in household dust over two seasons. No significant difference using Mann Whitney test.

The high Mtb prevalence via human shedding in dust was observed in TB case households with fifteen having especially high levels of shedding reaching 10^6 genomic equivalents g^{-1} (Figure 7.4A). In contrast to TB case households, Mtb shedding was negligible in control households with only two positives reaching 10^6

genomic equivalents g^{-1} (Figure 7.4B). It is worth noting that the Mtb positives in these two control individual households were in close proximity to TB case households with heavy Mtb shedding. These high loads of Mtb positives shedding across both households sets indicated considerable infection associated with human health conditions in this area. In addition the majority of Mtb shedding was similar between these two household sets but Mtb positives in Tb case household dust was more intense in the area (Figure 7.5).

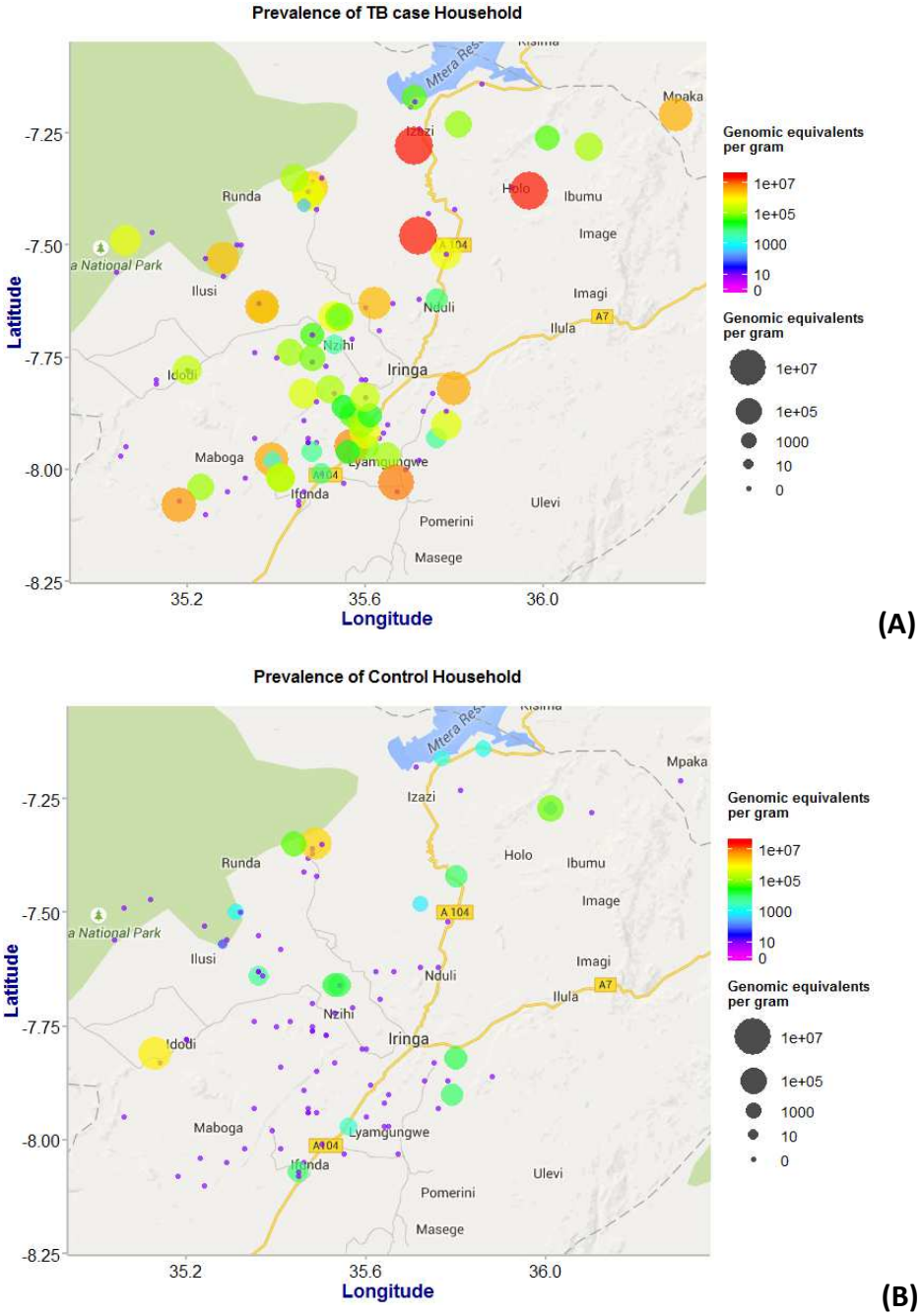


Figure 7.4 Mtb positives shedding in TB case households (A) and control households (B) in the study region.

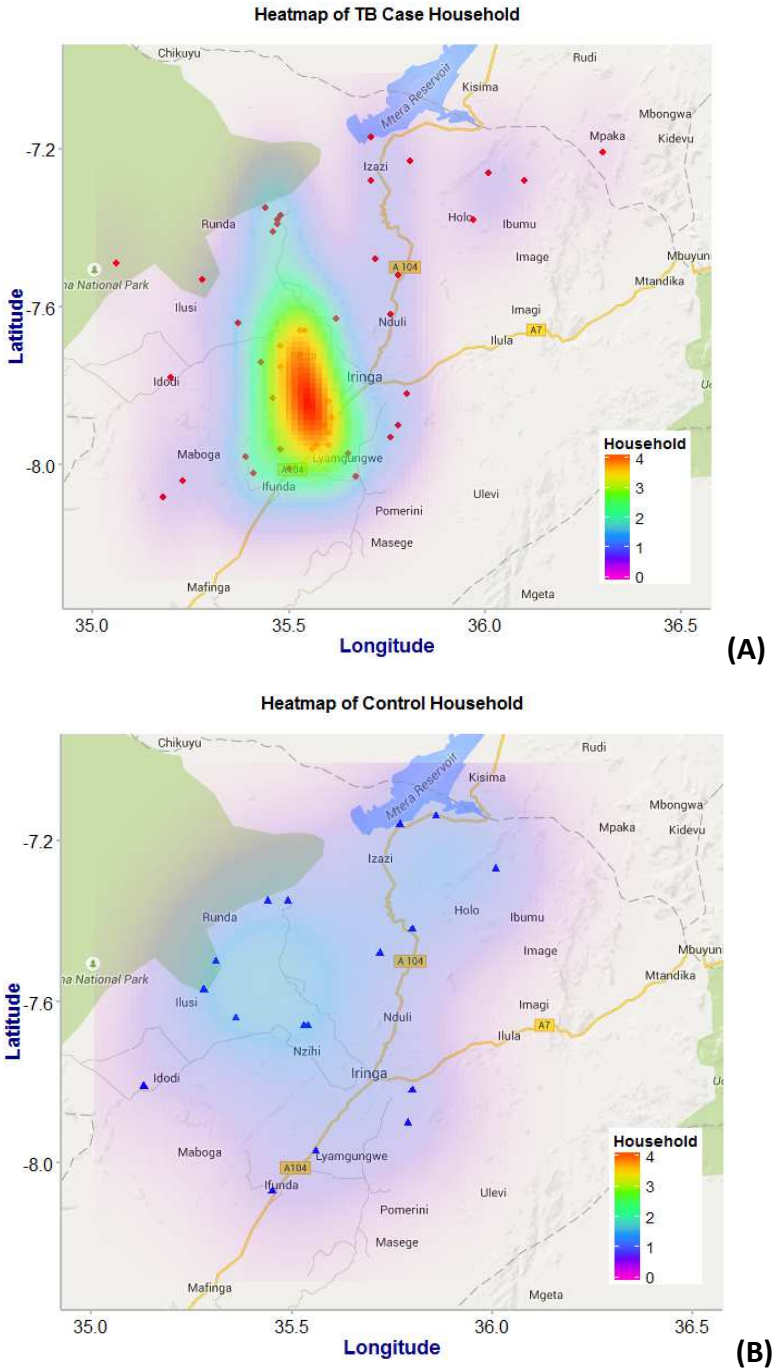


Figure 7.5 Density of Mtb positives shedding in TB case households (A) and Control households (B).

In total, 81.93 % of households were paired with TB case and control households but only in ten paired households were Mtb positives detected in soil (Figure 7.6 and Table 7.2A). There was no significant difference (P-value > 0.05) in terms of Mtb in

both household dusts samples sets using paired signed statistical analysis which compared Mtb contained in each TB case household with their matched control household (Table 7.2B)

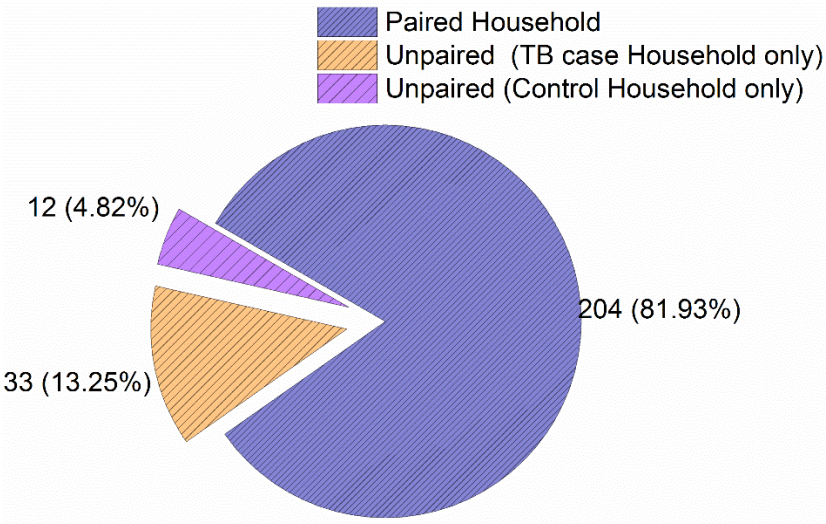


Figure 7.6 The proportion of paired and unpaired households.

Table 7.2 Comparison of Mtb testing results in paired household using McNemar's Test (A) and no different using paired signed test across two household sets (B).

		Control Households		
TB case Households		+	-	Total
	+	10	35	45
	-	12	45	57
	Total	22	80	102

(A)

(B)

Signed Test

Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
"TB Case Households"	10	7566.76407	42093.34068	455278.47168	1.77263E6	2.94292E7
"Control Households"	10	383.34238	2770.41332	15142.11698	31582.14996	2.13739E6

Test Statistics

S	Z	Prob>|S|
2	--	0.10938

Null Hypothesis: F(x) = G(y)
Alternative Hypothesis: F(x) <> G(y)
At the 0.05 level, the two distributions are NOT significantly different.

7.4. Discussion

The NTM were abundant in the environmental reservoirs especially in soil and water but Mtb was identified rarely from the environment compared to human TB patients (Falkinham 2009). The Mtb genotype was still identified from DNA in soil and water to compare with Mtb bacterial strains obtained from identified TB patients. Even the most dominant of Mtb genotypes in TB patients are inconsistent with their associated environmental genotypes (Velayati, Farnia et al. 2015). It is also worth noting that the survival of Mtb in soil was reported to extend to 12 months and Mtb still remains virulent to mice (Ghodbane, Mba Medie et al. 2014).

An NIH report revealed that TB patients in Tanzania shared the same local water supplies with non-TB patients because of the lack of public water supplies in this rural area but there were shared public wells in the village (Hopewell 2011). Water source contaminated by TB patients have been implied as the source of TB infection in Denmark (Steentoft, Wittendorf et al. 2006). Based on these publications, water and soil provide a potential risk factor for spreading TB disease within homesteads between TB patients and healthy humans in particular if they lived very close to their TB neighbour.

According to the preliminary data high Mtb positives were detected more frequently in the individual households with identified TB patients compared to control households which might be attributed to Mtb cells potentially carried from their TB neighbourhood or spread into air to nearby households. This situation increased the potential risk of human exposure to the TB patients if these patients cannot be isolated from communities and still share same water supplies and human shedding.

Bio-climatic conditions were evident as a strong influence on the mycobacteria community structure in soil (Khera 2012). However, these results are based on dust samples collected inside human dwellings, preventing high temperature and light irradiation effects thus increasing survival of Mtb in soil compared to exposure outside of homesteads. .

This study demonstrated that the prevalence and quantification of Mtb positives via TB case household soil were consistent with their matched control household if they both had Mtb pathogen. This evidence revealed that Mtb was capable of transfer from TB individual household to their close neighbour households.

In conclusion, this study revealed successfully that the reservoir of Mtb in TB case households compared to control households. These results indicate the potential transmission was established for a clinical human pathogen to its environment via shedding and on to susceptible individual in the neighbourhood. However the dominant Mtb strains should be strain typed and the whole genome SNP analysed to prove sources of Mtb strains from household dust. In addition diversity analysis will be applied to understand the impact of a single human mycobacterial pathogen species on the whole microbial community composition in dust. In contrast to Mtb, no Mb was detected in dust samples in all households consisting of TB case and control households.

A non-invasive test was established to prove that shedding into the environment is a diagnostic feature of ongoing disease. For example, we established 20 control households which give positive results for Mtb in dust. These households need to be checked now for inhabitants infectious status.

Chapter 8 Final discussion and conclusion

8.1 overview

This project aimed to explore diversity and distribution of SGM particularly two MTBC members, Mtb and Mb, in environmental samples in human communities with diverse activities, using effective molecular methods and NGS sequencing approaches. The environmental screening studies have revealed that diverse sample types may affect the survival of Mtb and Mb. The sequencing approaches illustrated the diversity of the mycobacteria community composition in the environment. This comprehensive study has demonstrated the high diversity of mycobacteria present in Tanzanian livestock faeces, soil and water. In conclusion, livestock faeces are a significant reservoir of mycobacteria species including Mb and Mtb and can be considered as a potential reservoir of infectious agents in the environment.

8.2. Prevalence and spatial study of Mtb and Mb

Many prokaryotes, including pathogenic bacteria, persist in the environment and cause human and animal disease upon exposure to high pathogen loads in the environment (Guan and Holley 2003). In addition several opportunistic mycobacterial species are known to persist under extreme environmental conditions (Primm, Lucero et al. 2004). Contamination of the environment from infected animals may be a potential infection route for pathogenic mycobacterial species, particularly in SGM infection transmission. The typical recognised transmission route of Mb is via aerosol or close contact to lesions (P 2011). But our research group has detected significant environmental contamination with Mb (Sweeney, Courtenay et al. 2007). Cattle

faeces are a potential reservoir of Mb infection and may spread bTB when cattle are eating a food source contaminated with infected faeces (Phillips, Foster et al. 2003).

The result of this study revealed that the majority of Mb and Mtb shedding in environmental samples is from both cattle and goat faeces, especially cattle faecal samples which had a higher proportion and higher load of Mb and Mtb positives. Although information on bTB infectious disease in cattle herds could not be obtained for this period, the results reflect high Mb prevalence in cattle herds in particular in pastoralist villages.

As previously mentioned in Chapter 4, Mb infected soil from boma soil was related to faecal shedding by positives with high levels of the same pathogens. This result is evidence that the pathogen is able to persist in soil when contaminated faeces decompose in mixtures of soil. Previous work in our research group identified that climatic conditions have a stronger influence on mycobacteria community structure than geographical proximity (Khera 2012). The work presented here show that the survival of Mtb and Mb in soil samples in this study was significantly correlated with soil moisture in the wet season here these bacteria would be less prone to environmental exposure to light irradiation effects. Mtb and Mb can possibly become airborne, spreading in the air and leading to infection, similarly to other SGM species, with MAP suggested as a potential bio-aerosol in MAP infection in a previous study (Eisenberg, Nielen et al. 2010).

Although few Mb positives were detected via contaminated water supplies, the river sampling locations were close to villages with high Mb prevalence. Therefore the

water environment may present a potential for the faecal-oral route of Mb transmission from water to livestock.

8.3. Diversity and abundance of mycobacteria

Pyrosequencing provided diversity analysis for SGM from diverse environmental samples compared to conventional molecular approaches such as qPCR. The structure of SGM communities was compared between sample types among pastoralist villages. Different profiles of *Mycobacterium* species were observed between samples types and different sequencing analysis methods were employed for analysis of the pyrosequencing dataset to evaluate which approach was suitable for diversity analysis.

QIIME is a conventional sequencing analysis platform with numerous algorithms for 16S rRNA metadata analysis (Kuczynski, Stombaugh et al. 2012). The species classification using 16S rRNA amplicons in QIIME still relies on sequence similarity to define OTU against a curated taxonomic database. However sequence similarity approaches neglected the less abundant but important bacterial species that had been clustered in favour of the most abundant sequences. In addition the difficulty of assigning taxonomic status to uncultured bacterial sequences limited species classification and diversity description (Zheng, Kramer et al. 2012).

Oligotyping was based on entropy analysis of variable sites in sequences to identify highly discriminatory positions (Eren, Maignien et al. 2013). This approach also reduced the proportion of unrecognised species from 12.9 % in QIIME to 2.98 % using Oligotyping when Oligotyping and QIIME both rely on BLAST as reference database for comparison of species classification. It suggests that Oligotyping has very efficient

noise removal using entropy analysis compared to application of OTU based sequence similarity threshold in QIIME.

Comparison of ranking distance based NMDs plot in QIIME, demonstrated all samples clustered together in QIIME analysis. In contrast to QIIME, Oligotyping results depicted strong differentiation between samples types. Both results are based on the same database but produce extremely different results in diversity analysis. It could be suggested that the different approaches of species identification including miss-match species identification or proportion of unknown species results in the difference in the diversity analysis. The SGM pre-test is another example which identified that the positives were not distinguished from negatives in QIIME as both sample types share high similarity in species abundance level but greater differentiation occurs in Oligotyping analysis.

The application of Oligotyping to this problem was used to cluster sequence by sequence entropy selection to avoid losing subtle and less dominant species and eliminate the effect of sequence noise. In addition Oligotyping eliminated a large proportion of unassigned bacterial species in the current study. Oligotyping identified two dominant SGM species, *M. asiaticum* and *M. intracellulare,* from Tanzanian environmental samples. The classification based on curated reference database was still applied for species identification in Oligotyping, however, the novel MED approach enabled reliable and sensitive sequence classification without comparing with databases and only separated according to differences in entropy node (Eren, Morrison et al. 2015).

The current study compared QIIME, Oligotyping and MED to perform environmental sample screening. The best separation of each environmental samples was from MED compared to other two algorithms. MED combined with statistical analysis also provided more accurate diversity analysis to delineate the relationship of bacterial communities between each environmental samples types. The bacterial diversity in livestock shedding faecal samples was significantly different from environmental samples, including boma soil. The MED results revealed an important aspect matched to the prevalence study of Mtb and Mb in terms of SGM diversity analysis, for example highly similar microbial diversity was detected between cattle and goat related samples as same as Oligotyping results shown. Another example associated with Mb potential transmission from water to cattle samples is observed in sequencing analysis, Oligotyping demonstrated similarity of SGM community structure between cattle and water related samples. These results implies the bacterial communities can be shared between livestock herds as well as between water and cattle.

8.4. NIH project implications of findings

The NIH project study also revealed the strong correlation of Mtb pathogens between TB patients and environmental soil associated with their households in agriculturalist villages. In contrast to high prevalence of Mtb from agriculturalist villages with identified TB patients, only some Mtb positives presented in the household dust of pastoralist villages without TB patients with only 4.4 % Mtb positives in the dry season and only 1.3 % in the wet season. Recent sequencing work revealed that Mtb and Mb are a minority of the SGM species in livestock faeces and other environmental

reservoirs, however, high prevalence of Mtb still presented in household dust particularly from households with TB patients. A recent publication from our research group indicated that animal faeces contained a high prevalence of Mb and as such represents a potential environmental reservoir of bTB, and possible transmission route (Hayley King 2015). These results imply that the proportion of pathogenic mycobacteria is variable, and can be detected when infected humans or animals shed into the environment.

8.5 Conclusion

This study concludes that the livestock faeces have the highest concentration of Mtb and Mb when compared to other different environmental samples. Mtb and Mb persist in the soil associated with contaminated faeces but this is dependent on moisture content in soil and light irradiation effects. Water is also a reservoir of bTB and thus a potential transmission route to livestock. Diversity analysis also shows high similarity in SGM species composition between water, sediment and cattle related faecal and soil samples. Although the majority of SGM species are not Mtb and Mb, they still represent a potential risk when humans or animals are exposed to a high concentrations in environmental reservoirs.

Oligotyping is better than QIIME in sequencing analysis including noise removal and species classification but still relies on annotated species reference databases. Oligotyping based MED optimised the differentiation of samples types and produce similar results without comparison to curated reference databases when compared to Oligotyping.

A strong correlation between Mtb prevalence in household dust and TB positive patient status was established in the NIH study. Soil in the household is considered as an environmental reservoir and a possible human TB transmission route.

8.6. Future work

The study aimed to characterise Mtb, Mb and other pathogenic SGM species in diversity and abundance analyses in the environment associated with human activities and livestock in Tanzania. However the genotypes of Mb and Mtb from infected human or livestock were not obtained to compare with Mtb and Mb strains detected from environmental sources. Genotyping analysis of strains from the environment, livestock, and human patients could indicate transmission pathways.

A SGM specific 16S rRNA modified MiSeq amplicon has been developed and can be used to screen environmental samples to compare composition of SGM between different variables. MiSeq data will help to correlate SGM diversity to seasonal change or other climatic factors such as precipitation in the environment. In addition other gene markers such as *rpoB* and *hsp65* should be used to confirm the diversity and prevalence of mycobacteria species particularly two prevalent opportunistic mycobacteria species, *Mycobacterium asiaticum* and *Mycobacterium intracellulare* These would provide evidence of a new aspect of NTM transmission route from the environment.

Reference

Ameni, G., M. Vordermeier, R. Firdessa, A. Aseffa, G. Hewinson, S. V. Gordon and S. Berg (2011). "Mycobacterium tuberculosis infection in grazing cattle in central Ethiopia." Vet J **188**(3): 359-361.

Anonymous (2009) "Quantiferon-TB Gold In-Tube Package Insert."

Barry, C. E., 3rd, R. E. Lee, K. Mdluli, A. E. Sampson, B. G. Schroeder, R. A. Slayden and Y. Yuan (1998). "Mycolic acids: structure, biosynthesis and physiological functions." Prog Lipid Res **37**(2-3): 143-179.

Berg, S., R. Firdessa, M. Habtamu, E. Gadisa, A. Mengistu, L. Yamuah, G. Ameni, M. Vordermeier, B. D. Robertson, N. H. Smith, H. Engers, D. Young, R. G. Hewinson, A. Aseffa and S. V. Gordon (2009). "The burden of mycobacterial disease in ethiopian cattle: implications for public health." PLoS One **4**(4): e5068.

Blacklock, Z. M., D. J. Dawson, D. W. Kane and D. McEvoy (1983). "Mycobacterium asiaticum as a potential pulmonary pathogen for humans. A clinical and bacteriologic review of five cases." Am Rev Respir Dis **127**(2): 241-244.

Boratyn, G. M., S. Datta and S. Datta (2006). "Biologically supervised hierarchical clustering algorithms for gene expression data." Conf Proc IEEE Eng Med Biol Soc **1**: 5515-5518.

Boulahbal, F., A. Benelmouffok and K. Brahimi (1978). "[Role of Mycobacterium tuberculosis in bovine tuberculosis]." Arch Inst Pasteur Alger **53**: 155-164.

Boxer, M. (2000). "Molecular techniques: divide or share." J Clin Pathol **53**(1): 19-21.

Brosch, R., S. V. Gordon, A. Pym, K. Eiglmeier, T. Garnier and S. T. Cole (2000). "Comparative genomics of the mycobacteria." Int J Med Microbiol **290**(2): 143-152.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Chaves, D., A. Sandoval, L. Rodriguez, J. C. Garcia, S. Restrepo and M. M. Zambrano (2010). "[Comparative analysis of six Mycobacterium tuberculosis complex genomes]." Biomedica **30**(1): 23-31.

Clarke, K. R. (1993). "Non-parametric multivariate analysis of changes in community structure." Australian Journal of Ecology **18**: 117-143.

Clarke, S. C. (2005). "Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications." Expert Rev Mol Diagn **5**(6): 947-953.

Corner, L. A., M. John, P. G. Bundesen and P. R. Wood (1988). "Identification of Mycobacterium bovis isolates using a monoclonal antibody." Vet Microbiol **18**(2): 191-196.

Cosivi, O., J. M. Grange, C. J. Daborn, M. C. Raviglione, T. Fujikura, D. Cousins, R. A. Robinson, H. F. Huchzermeyer, I. de Kantor and F. X. Meslin (1998). "Zoonotic tuberculosis due to Mycobacterium bovis in developing countries." Emerg Infect Dis **4**(1): 59-70.

Courtenay, O., L. A. Reilly, F. P. Sweeney, V. Hibberd, S. Bryan, A. Ul-Hassan, C. Newman, D. W. Macdonald, R. J. Delahay, G. J. Wilson and E. M. Wellington (2006). "Is Mycobacterium bovis in the environment important for the persistence of bovine tuberculosis?" Biol Lett **2**(3): 460-462.

Crawford, P. A., J. R. Crowley, N. Sambandam, B. D. Muegge, E. K. Costello, M. Hamady, R. Knight and J. I. Gordon (2009). "Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation." Proc Natl Acad Sci U S A **106**(27): 11276-11281.

Cremonesi, P., B. Castiglioni, G. Malferrari, I. Biunno, C. Vimercati, P. Moroni, S. Morandi and M. Luzzana (2006). "Technical note: Improved method for rapid DNA extraction of mastitis pathogens directly from milk." J Dairy Sci **89**(1): 163-169.

Daborn, C. J., J. M. Grange and R. R. Kazwala (1996). "The bovine tuberculosis cycle--an African perspective." Soc Appl Bacteriol Symp Ser **25**: 27S-32S.

de la Rua-Domenech, R., A. T. Goodchild, H. M. Vordermeier, R. G. Hewinson, K. H. Christiansen and R. S. Clifton-Hadley (2006). "Ante mortem diagnosis of tuberculosis in cattle: a review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques." Res Vet Sci **81**(2): 190-210.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edgar, R. C. (2013). "UPARSE: highly accurate OTU sequences from microbial amplicon reads." Nat Methods **10**(10): 996-998.

Eisenberg, S. W., M. Nielen, W. Santema, D. J. Houwers, D. Heederik and A. P. Koets (2010). "Detection of spatial and temporal spread of Mycobacterium avium subsp. paratuberculosis in the environment of a cattle farm through bio-aerosols." Vet Microbiol **143**(2-4): 284-292.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison and M. L. Sogin (2013). "Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data." Methods Ecol Evol **4**(12).

Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis and M. L. Sogin (2015). "Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences." ISME J **9**(4): 968-979.

Eren, A. M., M. L. Sogin, H. G. Morrison, J. H. Vineis, J. C. Fisher, R. J. Newton and S. L. McLellan (2015). "A single genus in the gut microbiome reflects host preference and specificity." ISME J **9**(1): 90-100.

Eren, A. M., M. Zozaya, C. M. Taylor, S. E. Dowd, D. H. Martin and M. J. Ferris (2011). "Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation." PLoS One **6**(10): e26732.

Etchechoury, I., G. E. Valencia, N. Morcillo, M. D. Sequeira, B. Imperiale, M. Lopez, K. Caimi, M. J. Zumarraga, A. Cataldi and M. I. Romano (2010). "Molecular typing of Mycobacterium bovis isolates in Argentina: first description of a person-to-person transmission case." Zoonoses Public Health **57**(6): 375-381.

Etter, E., P. Donado, F. Jori, A. Caron, F. Goutard and F. Roger (2006). "Risk analysis and bovine tuberculosis, a re-emerging zoonosis." Ann N Y Acad Sci **1081**: 61-73.

Eurosurveillance editorial, t. (2013). "WHO publishes Global tuberculosis report 2013." Euro Surveill **18**(43).

Euzeby, J. P. (1997). "List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet." Int J Syst Bacteriol **47**(2): 590-592.

Falkinham, J. O., 3rd (2009). "Surrounded by mycobacteria: nontuberculous mycobacteria in the human environment." J Appl Microbiol **107**(2): 356-367.

Falkinham, J. O., 3rd, C. D. Norton and M. W. LeChevallier (2001). "Factors influencing numbers of Mycobacterium avium, Mycobacterium intracellulare, and other Mycobacteria in drinking water distribution systems." Appl Environ Microbiol **67**(3): 1225-1231.

Floss, H. G. and T. W. Yu (2005). "Rifamycin-mode of action, resistance, and biosynthesis." Chem Rev **105**(2): 621-632.

Garnier, T., K. Eiglmeier, J. C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon and R. G. Hewinson (2003). "The complete genome sequence of Mycobacterium bovis." Proc Natl Acad Sci U S A **100**(13): 7877-7882.

Garrity, G. M. (2004). Taxonomic Outline of the Prokaryotes. Bergey's Manual of Systematic Bacteriology. J. A. B. T. G. L. G.M. Garrity. New York, Springer**:** 399.

Gengenbacher, M. and S. H. Kaufmann (2012). "Mycobacterium tuberculosis: success through dormancy." FEMS Microbiol Rev **36**(3): 514-532.

Gerald H. Mazurek, M. D., Margarita E. Villarino, M.D (2003). "Guidelines for Using the QuantiFERON®-TB Test for Diagnosing Latent Mycobacterium tuberculosis Infection." Recommendations and Reports **52(RR02)**: 15-18.

Gey van Pittius, N. C., S. L. Sampson, H. Lee, Y. Kim, P. D. van Helden and R. M. Warren (2006). "Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions." BMC Evol Biol **6**: 95.

Ghodbane, R., F. Mba Medie, H. Lepidi, C. Nappez and M. Drancourt (2014). "Long-term survival of tuberculosis complex mycobacteria in soil." Microbiology **160**(Pt 3): 496-501.

Gibson, J., S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs and M. Hajibabaei (2014). "Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics." Proc Natl Acad Sci U S A **111**(22): 8007-8012.

Gonzalez, A. and R. Knight (2012). "Advancing analytical algorithms and pipelines for billions of microbial sequences." Curr Opin Biotechnol **23**(1): 64-71.

Gowtage-Sequeira, S., A. Paterson, K. P. Lyashchenko, S. Lesellier and M. A. Chambers (2009). "Evaluation of the CervidTB STAT-PAK for the detection of Mycobacterium bovis infection in wild deer in Great Britain." Clin Vaccine Immunol **16**(10): 1449-1452.

GP, K. (1984). The mycobacteria: a sourcebook (Part A). United States of America, Marcel Dekker Inc.

Grant, I. R. and L. D. Stewart (2015). "Improved detection of Mycobacterium bovis in bovine tissues using immunomagnetic separation approaches." Methods Mol Biol **1247**: 153-161.

Greendyke, R., M. Rajagopalan, T. Parish and M. V. Madiraju (2002). "Conditional expression of Mycobacterium smegmatis dnaA, an essential DNA replication gene." Microbiology **148**(Pt 12): 3887-3900.

Griffiths, R. I., A. S. Whiteley, A. G. O'Donnell and M. J. Bailey (2000). "Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition." Appl Environ Microbiol **66**(12): 5488-5491.

Guan, T. Y. and R. A. Holley (2003). "Pathogen survival in swine manure environments and transmission of human enteric illness--a review." J Environ Qual **32**(2): 383-392.

Gutierrez, M. C., S. Brisse, R. Brosch, M. Fabre, B. Omais, M. Marmiesse, P. Supply and V. Vincent (2005). "Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis." PLoS Pathog **1**(1): e5.

Halevy, S., A. D. Cohen and N. Grossman (2005). "Clinical implications of in vitro drug-induced interferon gamma release from peripheral blood lymphocytes in cutaneous adverse drug reactions." J Am Acad Dermatol **52**(2): 254-261.

Halse, T. A., V. E. Escuyer and K. A. Musser (2011). "Evaluation of a single-tube multiplex real-time PCR for differentiation of members of the Mycobacterium tuberculosis complex in clinical specimens." J Clin Microbiol **49**(7): 2562-2567.

Harmsen, D., S. Dostal, A. Roth, S. Niemann, J. Rothganger, M. Sammeth, J. Albert, M. Frosch and E. Richter (2003). "RIDOM: comprehensive and public sequence database for identification of Mycobacterium species." BMC Infect Dis **3**: 26.

Havlir, D. V., H. Getahun, I. Sanne and P. Nunn (2008). "Opportunities and challenges for HIV care in overlapping HIV and TB epidemics." JAMA **300**(4): 423-430.

Hayley King, A. M., Phillip James, Emma Travis, David Porter, Yu-Jiun Hung, Jason Saywer, Jennifer Cork, Richard Delahay, William Gaze, Orin Courtenay, and Elizabeth Wellington (2015). "The variability and seasonality of the environmental reservoir of Mycobacterium bovis shed by wild European badgers." Nature scientific reports.

Helb, D., M. Jones, E. Story, C. Boehme, E. Wallace, K. Ho, J. Kop, M. R. Owens, R. Rodgers, P. Banada, H. Safi, R. Blakemore, N. T. Lan, E. C. Jones-Lopez, M. Levi, M. Burday, I. Ayakaka, R. D. Mugerwa, B. McMillan, E. Winn-Deen, L. Christel, P. Dailey, M. D. Perkins, D. H. Persing and D. Alland (2010). "Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology." J Clin Microbiol **48**(1): 229-237.

Heller, L. C., M. Jones and R. H. Widen (2008). "Comparison of DNA pyrosequencing with alternative methods for identification of mycobacteria." J Clin Microbiol **46**(6): 2092-2094.

Hert, D. G., C. P. Fredlake and A. E. Barron (2008). "Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods." Electrophoresis **29**(23): 4618-4626.

Hill, T. C., K. A. Walsh, J. A. Harris and B. F. Moffett (2003). "Using ecological diversity measures with bacterial communities." FEMS Microbiol Ecol **43**(1): 1-11.

Hopewell, P. (2011). "NIH ICIDR-Research Plan Final report." 90.

Hotelling., H. (1933). "Analysis of a complex of statistical variables into principal comJournal of Educational Psychology,ponents." **24**(6): 417-441,498-520.

Howard, S. T. and T. F. Byrd (2000). "The rapidly growing mycobacteria: saprophytes and parasites." Microbes Infect **2**(15): 1845-1853.

Huber, J. A., D. B. Mark Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield and M. L. Sogin (2007). "Microbial population structures in the deep marine biosphere." Science **318**(5847): 97-100.

Humblet, M. F., M. L. Boschiroli and C. Saegerman (2009). "Classification of worldwide bovine tuberculosis risk factors in cattle: a stratified approach." Vet Res **40**(5): 50.

Huse, S. M., D. M. Welch, H. G. Morrison and M. L. Sogin (2010). "Ironing out the wrinkles in the rare biosphere through improved OTU clustering." Environ Microbiol **12**(7): 1889-1898.

Jost, R. (2007). Milk and Dairy Products. Ullmann's Encyclopedia of Industrial Chemistry.

JS, Y. (2003). Molecular detection of Mycobacterium bovis in the environment. Ph. D, University of Warwick.

Kaevska, M., S. Lvoncik, J. Lamka, I. Pavlik and I. Slana (2014). "Spread of Mycobacterium avium subsp. paratuberculosis through soil and grass on a mouflon (Ovis aries) pasture." Curr Microbiol **69**(4): 495-500.

Karakousis, P. C., R. D. Moore and R. E. Chaisson (2004). "Mycobacterium avium complex in patients with HIV infection in the era of highly active antiretroviral therapy." Lancet Infect Dis **4**(9): 557-565.

Katale, B. Z., E. V. Mbugi, E. D. Karimuribo, J. D. Keyyu, S. Kendall, G. S. Kibiki, P. Godfrey-Faussett, A. L. Michel, R. R. Kazwala, P. van Helden and M. I. Matee (2013). "Prevalence and risk factors for infection of bovine tuberculosis in indigenous cattle in the Serengeti ecosystem, Tanzania." BMC Vet Res **9**: 267.

Katale, B. Z., E. V. Mbugi, S. Kendal, R. D. Fyumagwa, G. S. Kibiki, P. Godfrey-Faussett, J. D. Keyyu, P. Van Helden and M. I. Matee (2012). "Bovine tuberculosis at the human-livestock-wildlife interface: is it a public health problem in Tanzania? A review." Onderstepoort J Vet Res **79**(2): 463.

Kaul, K. L. (2001). "Molecular detection of Mycobacterium tuberculosis: impact on patient care." Clin Chem **47**(8): 1553-1558.

Kazwala, R. R., C. J. Daborn, J. M. Sharp, D. M. Kambarage, S. F. Jiwa and N. A. Mbembati (2001). "Isolation of Mycobacterium bovis from human cases of cervical adenitis in Tanzania: a cause for concern?" Int J Tuberc Lung Dis **5**(1): 87-91.

Khera, T. (2012). The diversity and distribution of Mycobacterium species in varying ecological and climatic environments. Doctor of Philosophy, University of Warwick.

Kim, H., S. H. Kim, T. S. Shim, M. N. Kim, G. H. Bai, Y. G. Park, S. H. Lee, G. T. Chae, C. Y. Cha, Y. H. Kook and B. J. Kim (2005). "Differentiation of Mycobacterium species by analysis of the heat-shock protein 65 gene (hsp65)." Int J Syst Evol Microbiol **55**(Pt 4): 1649-1656.

King, H. C., A. Murphy, P. James, E. Travis, D. Porter, J. Sawyer, J. Cork, R. J. Delahay, W. Gaze, O. Courtenay and E. M. Wellington (2015). "Performance of a Noninvasive Test for Detecting Mycobacterium bovis Shedding in European Badger (Meles meles) Populations." J Clin Microbiol **53**(7): 2316-2323.

Kolb, J., D. Hillemann, P. Mobius, J. Reetz, A. Lahiri, A. Lewin, S. Rusch-Gerdes and E. Richter (2014). "Genetic characterization of German Mycobacterium avium strains isolated from different hosts and specimens by multilocus sequence typing." Int J Med Microbiol **304**(8): 941-948.

Konstantinidis, K. T., A. Ramette and J. M. Tiedje (2006). "The bacterial species definition in the genomic era." Philos Trans R Soc Lond B Biol Sci **361**(1475): 1929-1940.

Kruskal., J. B. (1964). "Nonmetric multidimensional scaling: a numerical method." Psychometrika **29**: 115-129.

Kubica, G. P. (1984). The mycobacteria: a sourcebook (Part A). USA, Marcel Dekker Inc

Kuczynski, J., C. L. Lauber, W. A. Walters, L. W. Parfrey, J. C. Clemente, D. Gevers and R. Knight (2012). "Experimental and analytical tools for studying the human microbiome." Nat Rev Genet **13**(1): 47-58.

Kuczynski, J., J. Stombaugh, W. A. Walters, A. Gonzalez, J. G. Caporaso and R. Knight (2012). "Using QIIME to analyze 16S rRNA gene sequences from microbial communities." Curr Protoc Microbiol **Chapter 1**: Unit 1E 5.

Kumar, V. A., Abul K.; Fausto, Nelson; & Mitchell, Richard N. (2007). Robbins Basic Pathology (8th ed.), Saunders Elsevier.

Kunin, V., A. Engelbrektson, H. Ochman and P. Hugenholtz (2010). "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates." Environ Microbiol **12**(1): 118-123.

L.R. Bakken, J. D. v. E., J.T. Trevor, E.M.H. Wellington (Eds.) (1997). "Culturable and non-culturable bacteria in soil." Modern soil microbiology 47-61.

Lantz PG, M. M., Wadstro¨m T, Ra°dstro¨m P (1997). "Removal of PCR inhibitors from human faecal samples through the use of an aqueous two-phase system for sample preparation prior to PCR." J Microbiol Meth **28**: 159-167.

Larsson, T.-B. (2001). Ecological Bulletins, Biodiversity Evaluation Tools for European Forests, Wiley-Blackwell.

Leclerc, M. C., N. Haddad, R. Moreau and M. F. Thorel (2000). "Molecular characterization of environmental mycobacterium strains by PCR-restriction fragment length polymorphism of hsp65 and by sequencing of hsp65, and of 16S and ITS1 rDNA." Res Microbiol **151**(8): 629-638.

Lozupone, C., M. E. Lladser, D. Knights, J. Stombaugh and R. Knight (2011). "UniFrac: an effective distance metric for microbial community comparison." ISME J **5**(2): 169-172.

Lozupone, C. A., M. Hamady, S. T. Kelley and R. Knight (2007). "Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities." Appl Environ Microbiol **73**(5): 1576-1585.

Ma, Y., F. Pan and M. McNeil (2002). "Formation of dTDP-rhamnose is essential for growth of mycobacteria." J Bacteriol **184**(12): 3392-3395.

Magurran., A. E. (2004). Measuring Biological Diversity., Blackwell.

Malama, S., T. B. Johansen, J. B. Muma, S. Mwanza, B. Djonne and J. Godfroid (2014). "Isolation and molecular characterization of Mycobacterium bovis from Kafue lechwe (Kobus leche kafuensis) from Zambia." Trop Anim Health Prod **46**(1): 153-157.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

Marsh, S. (2007). "Pyrosequencing applications." Methods Mol Biol **373**: 15-24.

Marshall, H. M., R. Carter, M. J. Torbey, S. Minion, C. Tolson, H. E. Sidjabat, F. Huygens, M. Hargreaves and R. M. Thomson (2011). "Mycobacterium lentiflavum in drinking water supplies, Australia." Emerg Infect Dis **17**(3): 395-402.

Masson, A. M. and F. H. Prissick (1956). "Cervical lymphadenitis in children caused by chromogenic Mycobacteria." Can Med Assoc J **75**(10): 798-803.

Mazurek, G. H., J. Jereb, P. Lobue, M. F. Iademarco, B. Metchock, A. Vernon, N. C. f. H. I. V. S. T. D. Division of Tuberculosis Elimination, C. f. D. C. Tb Prevention and Prevention (2005). "Guidelines for using the QuantiFERON-TB Gold test for detecting Mycobacterium tuberculosis infection, United States." MMWR Recomm Rep **54**(RR-15): 49-55.

Mbugi, E. V., B. Z. Katale, S. Kendall, L. Good, G. S. Kibiki, J. D. Keyyu, P. Godfrey-Faussett, P. Van Helden and M. I. Matee (2012). "Tuberculosis cross-species transmission in Tanzania: towards a One-Health concept." Onderstepoort J Vet Res **79**(2): 501.

McLellan, S. L., R. J. Newton, J. L. Vandewalle, O. C. Shanks, S. M. Huse, A. M. Eren and M. L. Sogin (2013). "Sewage reflects the distribution of human faecal Lachnospiraceae." Environ Microbiol **15**(8): 2213-2227.

Munoz-Mendoza, M., N. Marreros, M. Boadella, C. Gortazar, S. Menendez, L. de Juan, J. Bezos, B. Romero, M. F. Copano, J. Amado, J. L. Saez, J. Mourelo and A. Balseiro (2013). "Wild boar tuberculosis in Iberian Atlantic Spain: a different picture from Mediterranean habitats." BMC Vet Res **9**: 176.

Mwikuma, G., G. Kwenda, B. M. Hang'ombe, E. Simulundu, T. Kaile, S. Nzala, S. Siziya and Y. Suzuki (2015). "Molecular identification of non-tuberculous mycobacteria isolated from clinical specimens in Zambia." Ann Clin Microbiol Antimicrob **14**: 1.

Neill, C. E. O. (2010). Antibiotic resistant staphylococci in the agricultural environment: reservoirs of resistance and infection. Doctor of Philosophy, University of Warwick.

Nelson, M. C., H. G. Morrison, J. Benjamino, S. L. Grim and J. Graf (2014). "Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys." PLoS One **9**(4): e94249.

Nilakanta, H., K. L. Drews, S. Firrell, M. A. Foulkes and K. A. Jablonski (2014). "A review of software for analyzing molecular sequences." BMC Res Notes **7**: 830.

Novais, R. C., S. Borsuk, O. A. Dellagostin and Y. R. Thorstenson (2008). "Molecular inversion probes for sensitive detection of Mycobacterium tuberculosis." J Microbiol Methods **72**(1): 60-66.

Nugent, G., J. Whitford, I. J. Yockney and M. L. Cross (2012). "Reduced spillover transmission of Mycobacterium bovis to feral pigs (Sus scofa) following population control of brushtail possums (Trichosurus vulpecula)." Epidemiol Infect **140**(6): 1036-1047.

Olive, D. M. and P. Bean (1999). "Principles and applications of methods for DNA-based typing of microbial organisms." J Clin Microbiol **37**(6): 1661-1669.

P, H. (2011). "NIH ICIDR-Research Plan Final report." **90**.

Palmer, M. V., W. R. Waters, T. C. Thacker, R. Greenwald, J. Esfandiari and K. P. Lyashchenko (2006). "Effects of different tuberculin skin-testing regimens on gamma interferon and antibody responses in cattle experimentally infected with Mycobacterium bovis." Clin Vaccine Immunol **13**(3): 387-394.

Park, S. H. and A. Bendelac (2000). "CD1-restricted T-cell responses and microbial infection." Nature **406**(6797): 788-792.

Phillips, C. J., C. R. Foster, P. A. Morris and R. Teverson (2003). "The transmission of Mycobacterium bovis infection to cattle." Res Vet Sci **74**(1): 1-15.

Pickup, R., Rhodes, G., & Saunders, J. R. (2004). Extraction of microbial DNA from aquatic sources In Freshwater. Molecular Microbial Ecology Manual 2nd ed, Kluwer Academic Publishers**:** 41-52.

Pontiroli, A., T. T. Khera, B. B. Oakley, S. Mason, S. E. Dowd, E. R. Travis, G. Erenso, A. Aseffa, O. Courtenay and E. M. Wellington (2013). "Prospecting environmental mycobacteria: combined molecular approaches reveal unprecedented diversity." PLoS One **8**(7): e68648.

Pontiroli, A., E. R. Travis, F. P. Sweeney, D. Porter, W. H. Gaze, S. Mason, V. Hibberd, J. Holden, O. Courtenay and E. M. Wellington (2011). "Pathogen quantitation in complex matrices: a multi-operator comparison of DNA extraction methods with a novel assessment of PCR inhibition." PLoS One **6**(3): e17916.

Prabakaran, P., E. Streaker, W. Chen and D. S. Dimitrov (2011). "454 antibody sequencing - error characterization and correction." BMC Res Notes **4**: 404.

Primm, T. P., C. A. Lucero and J. O. Falkinham, 3rd (2004). "Health impacts of environmental mycobacteria." Clin Microbiol Rev **17**(1): 98-106.

Quigley, L., O. O'Sullivan, T. P. Beresford, R. Paul Ross, G. F. Fitzgerald and P. D. Cotter (2012). "A comparison of methods used to extract bacterial DNA from raw milk and raw milk cheese." J Appl Microbiol **113**(1): 96-105.

Quince, C., A. Lanzen, R. J. Davenport and P. J. Turnbaugh (2011). "Removing noise from pyrosequenced amplicons." BMC Bioinformatics **12**: 38.

Raqib, R., J. Rahman, A. K. Kamaluddin, S. M. Kamal, F. A. Banu, S. Ahmed, Z. Rahim, P. K. Bardhan, J. Andersson and D. A. Sack (2003). "Rapid diagnosis of active tuberculosis by detecting antibodies from lymphocyte secretions." J Infect Dis **188**(3): 364-370.

Reddington, K., J. O'Grady, S. Dorai-Raj, M. Maher, D. van Soolingen and T. Barry (2011). "Novel multiplex real-time PCR diagnostic assay for identification and differentiation of Mycobacterium tuberculosis, Mycobacterium canettii, and Mycobacterium tuberculosis complex strains." J Clin Microbiol **49**(2): 651-657.

Ronaghi, M., M. Uhlen and P. Nyren (1998). "A sequencing method based on real-time pyrophosphate." Science **281**(5375): 363, 365.

Rose, M. V., G. Kimaro, T. N. Nissen, I. Kroidl, M. Hoelscher, I. C. Bygbjerg, S. G. Mfinanga and P. Ravn (2012). "QuantiFERON(R)-TB gold in-tube performance for diagnosing active tuberculosis in children and adults in a high burden setting." PLoS One **7**(7): e37851.

Roug, A., A. Perez, J. A. Mazet, D. L. Clifford, E. VanWormer, G. Paul, R. R. Kazwala and W. A. Smith (2014). "Comparison of intervention methods for reducing human exposure to Mycobacterium bovis through milk in pastoralist households of Tanzania." Prev Vet Med **115**(3-4): 157-165.

Shinnick, T. M. and R. C. Good (1994). "Mycobacterial taxonomy." Eur J Clin Microbiol Infect Dis **13**(11): 884-901.

Shitaye JE, T. W., Pavlik I (2007). "Bovine tuberculosis infection in animal and human populations in Ethiopia: a review." Veterinarni Medicina **52**(8): 317-332.

Smith, D. P. and K. G. Peay (2014). "Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing." PLoS One **9**(2): e90234.

Stackebrandt, E., and J. Ebers. (2006). "Taxonomic parameters revisited: tarnished gold standards." Microbiol. Today **33**: 152-155.

Steentoft, J., J. Wittendorf and J. R. Andersen (2006). "[Tuberculosis and water pipes as source of infection]." Ugeskr Laeger **168**(9): 904-907.

Stewart, L. D., J. McNair, L. McCallan, S. Thompson, L. A. Kulakov and I. R. Grant (2012). "Production and evaluation of antibodies and phage display-derived peptide ligands for immunomagnetic separation of Mycobacterium bovis." J Clin Microbiol **50**(5): 1598-1605.

Sulieman, M. S. and M. E. Hamid (2002). "Identification of acid fast bacteria from caseous lesions in cattle in Sudan." J Vet Med B Infect Dis Vet Public Health **49**(9): 415-418.

Support, I. T. (2014). 16S Metagenomic Sequencing Library Preparation. I. T. Support.

Sweeney, F. P., O. Courtenay, V. Hibberd, R. G. Hewinson, L. A. Reilly, W. H. Gaze and E. M. Wellington (2007). "Environmental monitoring of Mycobacterium bovis in badger feces and badger sett soil by real-time PCR, as confirmed by immunofluorescence, immunocapture, and cultivation." Appl Environ Microbiol **73**(22): 7471-7473.

Sweeney, F. P., O. Courtenay, A. Ul-Hassan, V. Hibberd, L. A. Reilly and E. M. Wellington (2006). "Immunomagnetic recovery of Mycobacterium bovis from naturally infected environmental samples." Lett Appl Microbiol **43**(4): 364-369.

Taib, N., J. F. Mangot, I. Domaizon, G. Bronner and D. Debroas (2013). "Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity." PLoS One **8**(3): e58950.

Tanzania, N. T. a. L. P. (2006). " Annual Report for 2006." Dar es Salaam: Ministry of Health Tanzania.

Tedersoo, L., R. H. Nilsson, K. Abarenkov, T. Jairus, A. Sadam, I. Saar, M. Bahram, E. Bechem, G. Chuyong and U. Koljalg (2010). "454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases." New Phytol **188**(1): 291-301.

Thomas, T., J. Gilbert and F. Meyer (2012). "Metagenomics - a guide from sampling to data analysis." Microb Inform Exp **2**(1): 3.

Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel and M. F. Polz (2005). "Genotypic diversity within a natural coastal bacterioplankton population." Science **307**(5713): 1311-1313.

Torsvik, V., J. Goksoyr and F. L. Daae (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-787.

Torsvik, V., L. Ovreas and T. F. Thingstad (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science **296**(5570): 1064-1066.

Tortoli, E. (2003). "Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s." Clin Microbiol Rev **16**(2): 319-354.

Tortoli, E. (2012). "Phylogeny of the genus Mycobacterium: many doubts, few certainties." Infect Genet Evol **12**(4): 827-831.

Travis, E. R., W. H. Gaze, A. Pontiroli, F. P. Sweeney, D. Porter, S. Mason, M. J. Keeling, R. M. Jones, J. Sawyer, A. Aranaz, E. C. Rizaldos, J. Cork, R. J. Delahay, G. J. Wilson, R. G. Hewinson, O. Courtenay and E. M. Wellington (2011). "An inter-laboratory validation of a real time PCR assay to measure host excretion of bacterial pathogens, particularly of Mycobacterium bovis." PLoS One **6**(11): e27369.

V. Kunin, A. C., A. Lapidus, K. Mavromatis, and P. Hugenholtz. (2008). "A bioinformaticians guide to metagenomics." Microbiology and molecular biology reviews MMBR **72**(4): 557-578.

van Ingen, J., Z. Rahim, A. Mulder, M. J. Boeree, R. Simeone, R. Brosch and D. van Soolingen (2012). "Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies." Emerg Infect Dis **18**(4): 653-655.

van Vliet, P. C., J. W. Reijs, J. Bloem, J. Dijkstra and R. G. de Goede (2007). "Effects of cow diet on the microbial community and organic matter and nitrogen content of feces." J Dairy Sci **90**(11): 5146-5158.

Velayati, A. A., P. Farnia, M. Mozafari, D. Malekshahian, A. M. Farahbod, S. Seif, S. Rahideh and M. Mirsaeidi (2015). "Identification and genotyping of Mycobacterium tuberculosis isolated from water and soil samples of a metropolitan city." Chest **147**(4): 1094-1102.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Appl Environ Microbiol **73**(16): 5261-5267.

Webster, N. S., M. W. Taylor, F. Behnam, S. Lucker, T. Rattei, S. Whalan, M. Horn and M. Wagner (2010). "Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts." Environ Microbiol **12**(8): 2070-2082.

Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California." Ecological Monographs **30**: 279-338.

Williamson, D. A., I. Basu, J. Bower, J. T. Freeman, G. Henderson and S. A. Roberts (2012). "An evaluation of the Xpert MTB/RIF assay and detection of false-positive rifampicin resistance in Mycobacterium tuberculosis." Diagn Microbiol Infect Dis **74**(2): 207-209.

Wilson, I. G. (1997). "Inhibition and facilitation of nucleic acid amplification." Appl Environ Microbiol **63**(10): 3741-3751.

Wipat, A., M. H. Wellington and V. A. Saunders (1994). "Monoclonal antibodies for Streptomyces lividans and their use for immunomagnetic capture of spores from soil." Microbiology **140 (Pt 8)**: 2067-2076.

Young, J. S., E. Gormley and E. M. Wellington (2005). "Molecular detection of Mycobacterium bovis and Mycobacterium bovis BCG (Pasteur) in soil." Appl Environ Microbiol **71**(4): 1946-1952.

Zanini, M. S., E. C. Moreira, M. T. Lopes, P. Mota and C. E. Salas (1998). "Detection of Mycobacterium bovis in milk by polymerase chain reaction." Zentralbl Veterinarmed B **45**(8): 473-479.

Zheng, Z., S. Kramer and B. Schmidt (2012). "DySC: software for greedy clustering of 16S rRNA reads." Bioinformatics **28**(16): 2182-2183.

Zumla, A., A. George, V. Sharma, R. H. Herbert, I. Baroness Masham of, A. Oxley and M. Oliver (2015). "The WHO 2014 global tuberculosis report--further to go." Lancet Glob Health **3**(1): e10-12.

Appendix

Appendix 1 supplemental map

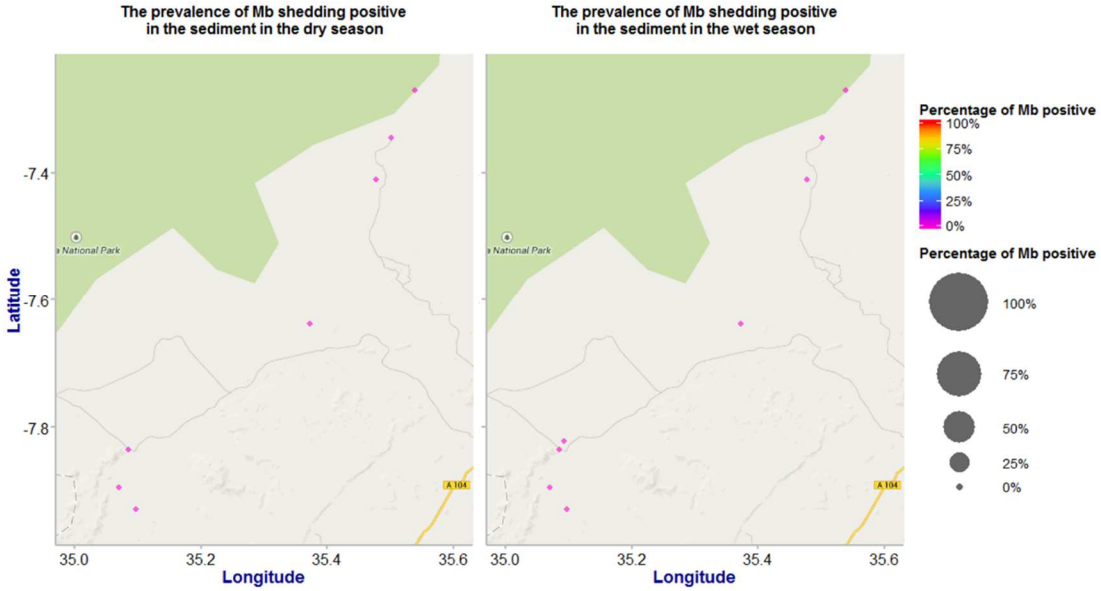


Figure A.1.1. No prevalence of Mb incidence in the sediment in the water sampling sites in each season; scale shown on the right indicates the percentage of Mb positive results.

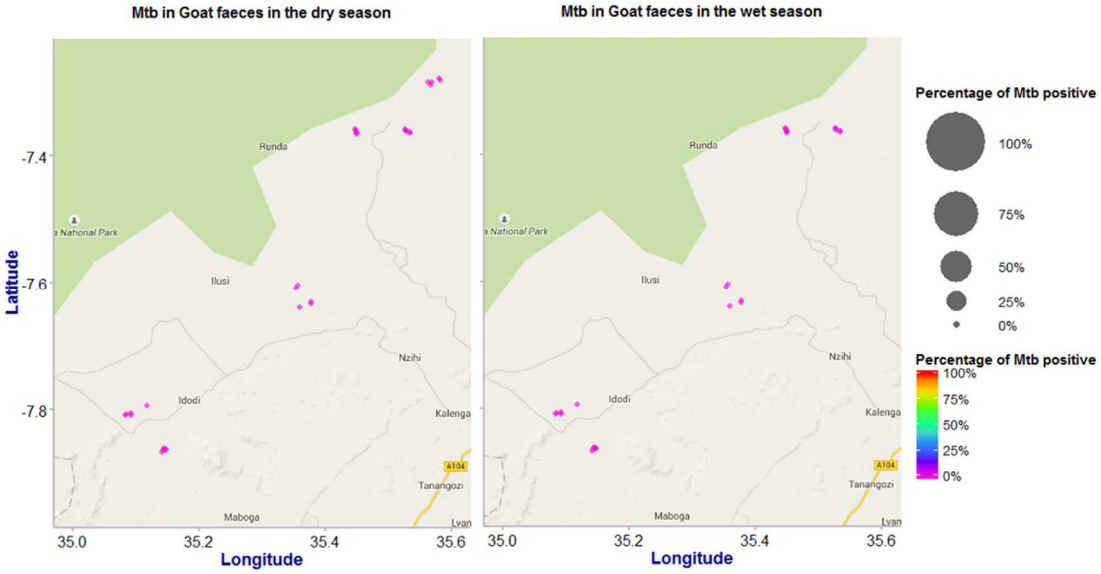


Figure A.1.2. Prevalence of Mtb incidence in the goat faeces in the six villages in each season; scale shown on the right indicates the percentage of Mtb positive results.

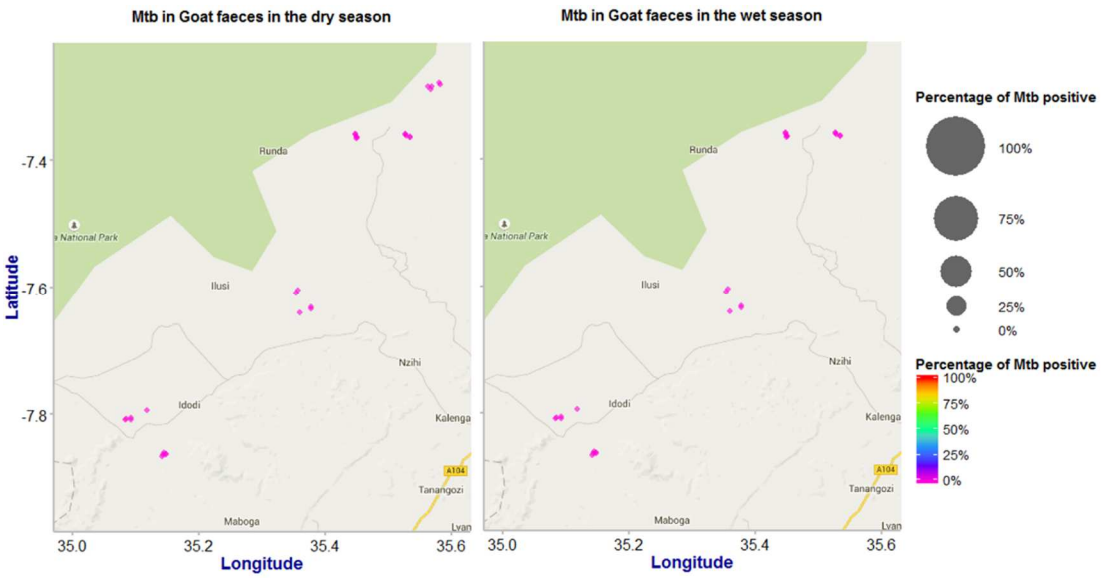


Figure A.1.3. Prevalence of Mtb incidence in the goat boma soil in the six villages in each season; scale shown on the right indicates the percentage of Mtb positive results.

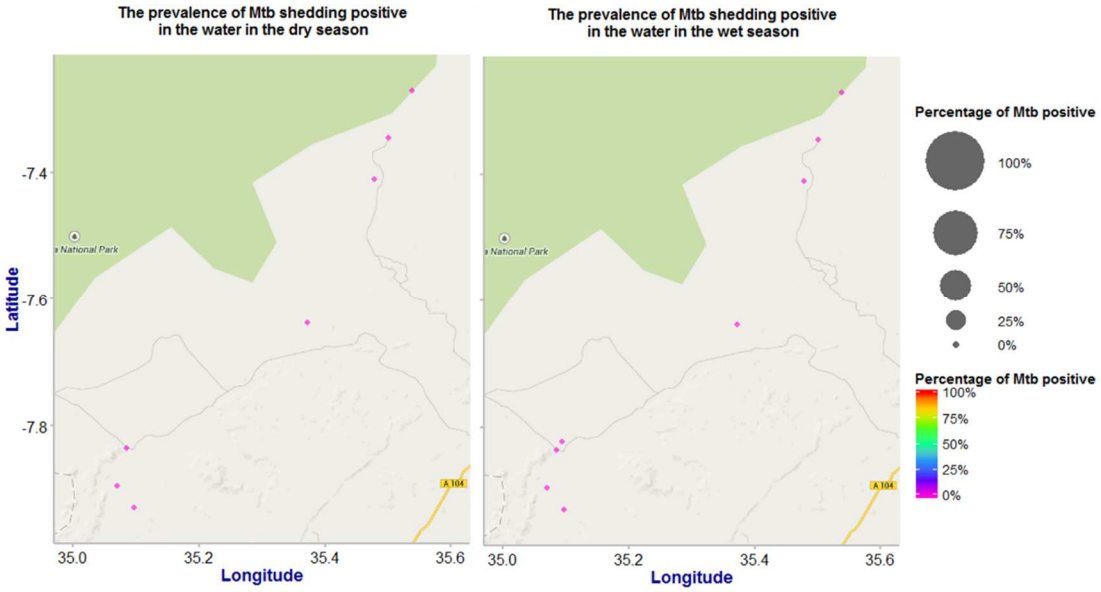


Figure A.1.4. Prevalence of Mtb incidence in the water samples in the eight water sampling sites in each season; scale shown on the right indicates the percentage of Mtb positive results.

Appendix 2: Miseq and Uparse analysis

A2.1. Samples preparation for MiSeq

A2.1.1. MiSeq: Modified primer design

The specific amplicon primer for Illumina MiSeq (ILLUMINA CAMBRIDGE LTD, Uttlesford Essex, UK) used in this study was obtained using APTK-MiSeq primers (Table 2.6). The APTK-MiSeq primer targeted specifically SGM species. The Illumina overhang adaptors sequence (Support 2014) (Figure A2.1) was prior to our specific primers in order to construct functional MiSeq primer. There were two amplicon PCR, two PCR clean-up and normalisation steps before sending to MiSeq analysis.

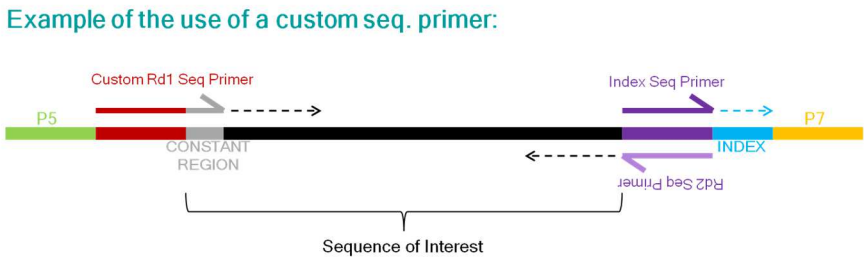


Figure A2.1. The example of amplicon construction consisted of the index seq Primers, P5, P7 region (adapter), custom primers and target sequence for MiSeq sequencing (Support 2014)

A2.1.2. MiSeq: First PCR

The highly conserved, the helix 18 insertion V3 region of the 16S rRNA (Figure 2.7) (Tortoli 2003) was used to target SGM species with primer pair APTK. The barcode sequence was introduced in the APTK primer for MiSeq. Optimism of APTK-MiSeq primer was critical due to APTK-MiSeq primer caused melt temperature to change compared to original. The optimal PCR reactions were conducted in a 50 µl reagent including 10 µl MyFi Reaction Buffer (Scientific Laboratory Supplies Ltd, Nottingham, UK), 1.2 µl MyFi DNA Polymerase (Scientific Laboratory Supplies Ltd, UK), 10 µl of

selected environmental DNA samples, 700 nM of APTK Forward-MiSeq and Reverse-MiSeq primers in final concentration and made up to total 50 µl with sterilised molecular grade water (Sigma-Aldrich Company Ltd, Dorset, UK).

The mixture reagent was loaded on to the MicroAmp 96-Well Reaction Plates and sealed with the X100 Adhesive PCR Film Polyester. 1 min centrifuge at 1250 RPM after plate setting-up and then placed into Bio-Rad T100™ Thermal Cycler PCR machine (Bio-Rad Laboratories Ltd., Hemel Hempstead, USA). This first optimal PCR programme for APTK-MiSeq was 95.0 °C for 4 min, followed by 45 cycles of 95.0 °C for 45 sec, 52.0 °C for 45 sec, 72.0 °C for 45 sec and a single extension step of 72.0 °C for 4 min.

A2.1.3. MiSeq: first clean-up

The first clean-up step followed by first PCR used 30 µl AMPure XP magnetic beads (Scientific Laboratory Supplies Ltd., Nottingham, UK) on DNA capture. These magnetic beads were attached by Magnetic Stand-96 device (Thermo Fisher Scientific, Leicestershire, UK) and was washed gently twice with 200 µl 80% fresh prepared ethanol. These steps were for purification of template amplicon and discrimination of redundant free primer and primer dimer species. The DNA products subsequently were eluted from magnetic beads with 15 µl 10 mM Tris pH 8.5 buffer for second PCR.

A2.1.4. MiSeq: index primer PCR

The indices and Illumina adaptor was attached to amplicon in the PCR using the Nextera XT Index Kit (ILLUMINA CAMBRIDGE LTD, Uttlesford Essex, UK). There were two different sets of index primer in the Nextera XT index kit as Index Primer 1 (N701 to N712) and Primer 2 (S517, S502 to S508). A set of index Primer 1 were arranged

horizontally and aligned with columns 1 through 12 on PCR plate and 5 µl of each primer 1 was loaded eight times separately from each row A to H. Another set of index Primer 2 were arranged vertically and aligned with row A through H on PCR plate and 5 µl of each primer 2 was loaded twelve times separately from each column 1 to 12. 6 µl MyFi Reaction Buffer (Scientific Laboratory Supplies Ltd, Nottingham, UK), 0.72 µl MyFi DNA Polymerase (Scientific Laboratory Supplies Ltd, UK), 10 µl template DNA were loaded and made up to total 30 µl with sterilised molecular grade water (Sigma-Aldrich Company Ltd, Dorset, UK). The mixture was loaded on to MicroAmp 96-Well Reaction Plates and sealed with X100 Adhesive PCR Film Polyester. The plate was spun down with 1 min at 1250 RPM and placed into Bio-Rad T100™ Thermal Cycler PCR machine with PCR programme 95.0 °C for 3 min, followed by 12 cycles of 95.0 °C for 30 sec, 55.0 °C for 30 sec, 72.0 °C for 30 sec and a single extension step of 72.0 °C for 5 min and hold at 4.0 °C to the end of programme.

A2.1.5. MiSeq: second clean-up

The second clean-up step was followed by index PCR similar to the first clean-up and the volume of the mixtures was adjusted. The 30µl AMPure XP magnetic beads in the first clean-up procedure was increased to 50 µl. The 15 µl 10 mM Tris pH 8.5 buffer was increased to 20 µl in the last step.

A2.1.6. MiSeq: Normalisation and Pool

The Qubit® dsDNA BR Assay Kit (Thermo Fisher Scientific, Leicestershire, UK) was employed for DNA quantification using with Qubit® 2.0 Fluorimeter (Thermo Fisher Scientific, Leicestershire, UK) following manufacturer's instructions. The 190 µl working solution was prepared by diluting the Qubit® dsDNA BR Reagent 1:200 in

Qubit® dsDNA BR Buffer. The 10 µl of each DNA product was mixed with the working solution and the mixture was tested with Qubit® 2.0 Fluorimeter. The DNA concentration was then gained and shown with unit, ng/µl, and the equation A2.1 as presented below was supplied to convert ng/µl to nM for calculation and sample pooling.

$$\frac{(\text{concentration in ng/µl})}{(660\ \text{g/mol} \times \text{average library size})} \times 10^6 = \text{concentration in nM}$$

Equation A2.1.

The DNA samples were diluted to 4 nM using 10 mM Tris pH 8.5 buffer and 5 µl aliquot of diluted DNA were pooled together for one MiSeq run.

A2.2.1. UPARSE analysis

UPARSE analysis was published in 2013 for constructing *de novo* OTU of sequence reads than achieve high accuracy in biological sequence recovery (Edgar 2013). The single programme of unique USEARCH with diverse algorithms was compared to QIIME which has a numbers of different clustering programmes.

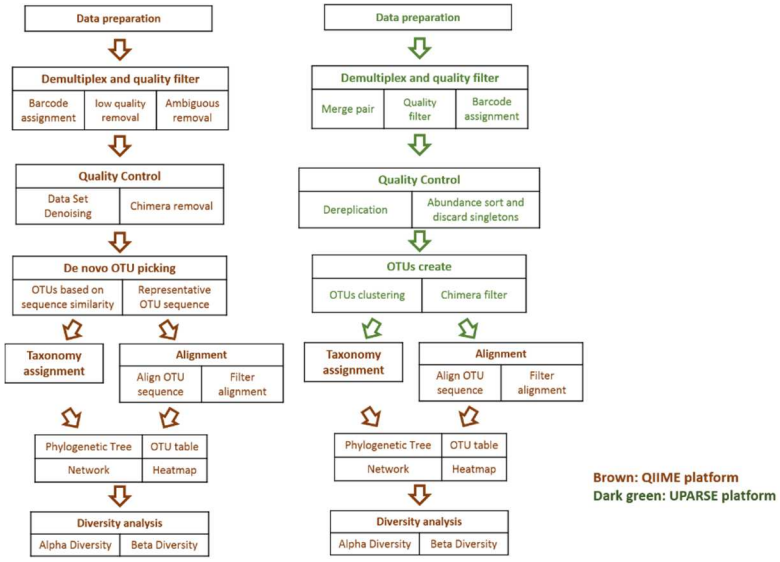


Figure A2.2. The comparison between QIIME and UPARSE procedure on analysis sequencing data.

A2.2.2. UPARSE: Preparation of raw Illumina data

The raw illumina data FASTQ file was a text-based format and was not required to convert to any format for downstream analysis. Each sample; however, had their own FASTQ file, which was different from SFF file, only one file included all data together. These FASTQ files were merged together to one file before downstream analysis

A2.2.3. UPARSE: Demultiplex and quality filter

There were two files for each sample set consisting of one file collected amplicon sequence from forward primer and another from reverse primer. The UPARSE script, *fastq_mergepairs* was employed to merge forward and reverse files together for sequence check and reduce the proportion of sequence errors. This step was capable of improving quality by comparison of reads from forward and reverse terminal while merging together.

Table A2.1. Description of Q score was represented and labelled by ASCII characters. The *P* was the possibility of an error.

Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

Phred or quality score (Q-score) was an integer value representing the estimated probability of an error occurring on each base call in one sequence and equation A2.2 described below the Q-score (Table A2.1):

$$Q = -10 \log_{10}(P)$$

Equation A2.2.

Q is quality and *P* probability of read error, for instance, *Q* as 3 and *P* was 0.5, meaning that there was 50% chance of an error. The expected value (E value) was an accumulation of Q score for each single nucleotide in a read. Small E value means high quality of this read and the E value was estimated to establish the prediction of quality level in a MiSeq run.

Amplicon reads was filtered based on E value using USEARCH method with UPARSE script, *fastq_filter*, with parameter, *fastq_maxee 1*, designed to discard the E-value > 1 in a run because the most probable number of errors of a filtered read is zero. Though of course, we expected some of them to have one or more errors. New barcodes were assigned to each sample for identification before pooling all sample data together (Linux terminal commands *sed* then *cat*).

A2.9.2.4. UPARSE: Quality Control

The first aim of quality control was to reserve the full-length reads to implement in an algorithm. The UPARSE script, *derep_fulllength*, was utilised to discriminate incomplete sequence such as prefixed and substrings sequence.

The second aim of quality control was to reduce the number of singletons. The singletons in a dataset normally was caused by random sequence errors distributed and reproduced by chance. These singletons was excluded without interfere of downstream analysis using UPARSE script, *sortbysize*.

A2.9.2.5. UPARSE: De novo OTU picking

The UPARSE script, *cluster_OTU*, was run to create OTU with abundance-sorted reads and generated representative sequence for each OTU.

Chimera sequences was filtered and eliminated after OTU creation using UCHIME with script, *uchime_ref*, with reference-based chimera filter as required. The reference used was *RDP_Gold* from the Ribosomal Database Project (http://rdp.cme.msu.edu/index.jsp) with > 10,000 reference sequences for 16S gene analysis.

Downstream analysis often required OTU table containing OTU representative sequence, a matrix that gave a read number to each sample and assigned to each OTU. The first step of creating OTU table was to match the reads back to OTU using the UPARSE script, *usearch_global*, and created OTU table with running python script, *uc2otutab.py*.

A2.9.2.6. UPARSE: Convert of UPARSE file into downstream QIIME analysis

The downstream analysis after OTU table was created and removal of chimera was imported into QIIME pipelines for taxonomy assignment, alignment and diversity analysis. The different format between UPARSE and QIIME was required to convert from OTU-table produced by UPARSE platform to OTU-table.biom used on QIIME. This step was accomplished using converting script, *biom convert*. The rest of downstream analysis and python script used were the same as QIIME.

Appendix 3: QIIME, Oligotyping and MED commands

For 454 pyrosequencing Fasta format data using QIIME

Denoising:

$ split_libraries.py -f 454allsample.fna -q 454allsample.qual -o 454allsample_demultiplexed/ -b 0 -m 454allsample_mapping.txt -M 1 -n 1000000 -l 400 -k

Re-integrating the denoised data into QIIME:

$ inflate_denoiser_output.py -c centroids.fna -s singletons.fna -f 454allsample.fna -d 454allsample_mapping.txt -o 454allsample_seqs.fna

Chimera checking and removel:

$ identify_chimeric_seqs.py -m blast_fragments -i 454sample.fna -t 13_8-release/taxonomy/99_otu_taxonomy.txt -r gg_13_8-release/rep_set/99_otus.fasta -o 454allsample_chimeric_seqs.txt

$ filter_fasta.py -f 454sample.fna -o non_chimeric_ aligned.fasta -s 454allsample_chimeric_seqs.txt -n

OTU clustering:

$ pick_otus.py -i non_chimeric_aligned.fasta -o picked_otus_default -s 0.97

$ pick_otus.py -i non_chimeric_aligned.fasta -o picked_otus_default -s 0.80

Pick representative set of sequences:

$ pick_rep_set.py -i seqs_otus.txt -f seqs.fna -o rep_set.fna

Aligning the representative OTU:

$ align_seqs.py -i $PWD/ rep_set.fna -t $PWD/core_set_aligned.fasta.imputed -o $PWD/pynast_aligned/

$ align_seqs.py -i rep_set.fna -t $PWD/gg_97_otus_6oct2010_aligned.fasta -o $PWD/pynast_aligned/

Assign taxonomy:

$ assign_taxonomy.py -t gg_13_8-release/taxonomy/99_otu_taxonomy.txt -r gg_13_8-release/rep_set/99_otus.fasta -i otus.fa -o assigned_taxonomy

$ assign_taxonomy.py -t gg_13_8-release/taxonomy/99_otu_taxonomy.txt -r gg_13_8-release/rep_set/99_otus.fasta -i otus.fa -o assigned_taxonomy -m blast

$ assign_taxonomy.py -t gg_13_8-release/taxonomy/99_otu_taxonomy.txt -r gg_13_8-release/rep_set/99_otus.fasta -i otus.fa -o assigned_taxonomy -m rdp

Building a phylogenetic tree:

$ make_phylogeny.py -i $PWD/aligned.fasta -o $PWD/rep_phylo.tre

Make OTU Heatmap (optional):

$ make_otu_heatmap_html.py -i otu_table.biom -o OTU_Heatmap/

Summarize Communities by Taxonomic Composition:

$ summarize_taxa_through_plots.py -i otu_table.biom -o wf_taxa_summary -m Fasting_Map.txt

$ summarize_taxa_through_plots.py -i otu_table.biom -o wf_taxa_summary -m Fasting_Map.txt -c Sample

$ summarize_taxa_through_plots.py -i otu_table.biom -o wf_taxa_summary -m Fasting_Map.txt -c Species

$ summarize_taxa_through_plots.py -i otu_table.biom -o wf_taxa_summary -m Fasting_Map.txt -c Region

$ summarize_taxa_through_plots.py -i otu_table.biom -o wf_taxa_summary -m Fasting_Map.txt -c Desceiption

Alpha Diversity:

$ echo "alpha_diversity:metrics shannon,PD_whole_tree,chao1,observed_species" > alpha_params.txt

$ alpha_rarefaction.py -i otu_table.biom -m Fasting_Map.txt -o wf_aldiv/ -p alpha_params.txt -t rep_set.tre

$ alpha_rarefaction.py -i otu_table.biom -m Fasting_Map.txt -o wf_aldiv/ -p alpha_params.txt -t rep_set.tre -e 300

$ compare_alpha_diversity.py -i shannon_tree.txt -m mapping.txt -c Sample-d 300 -o Sample_PD300

$ compare_alpha_diversity.py -i shannon_tree.txt -m mapping.txt -c Species-d 300 -o Species_PD300

$ compare_alpha_diversity.py -i shannon_tree.txt -m mapping.txt -c Region-d 300 -o Region_PD300

$ compare_alpha_diversity.py -i shannon_tree.txt -m mapping.txt -c Description-d 300 -o SGMtest_PD300

Beta Diversity:

$ beta_diversity_through_plots.py -i otu_table.biom -m Fasting_Map.txt -o wf_bdiv/ -t rep_set.tre

$ beta_diversity_through_plots.py -i otu_table.biom -m Fasting_Map.txt -o wf_bdiv/ -t rep_set.tre -e 300

$ compare_categories.py --method dbrda -i unweighted_unifrac_dm.txt -m Fasting_Map.txt -c Sample -o dbrda_Sample_unweight_out

$ compare_categories.py --method dbrda -i weighted_unifrac_dm.txt -m Fasting_Map.txt -c Sample -o dbrda_Sample_weight_out

$ compare_categories.py --method permdisp -i unweighted_unifrac_dm.txt -m Fasting_Map.txt -c Sample -o permdisp_Sample_unweight_out

$ compare_categories.py --method permdisp -i weighted_unifrac_dm.txt -m Fasting_Map.txt -c Sample -o permdisp_Sample_weight_out

$ compare_categories.py --method anosim -i unweighted_SvsW.txt -m mapping_allsample.txt -c Sample -o anosim_unw_SvsW

$ compare_categories.py --method anosim -i weighted_SvsW.txt -m mapping_allsample.txt -c Sample -o anosim_wei_SvsW

For MiSeq Fastq format data using Uparse:

Demultiplex:

- **Merge pair**

$ usearch -fastq_mergepairs sample1_R1.fastq -reverse sample1_R2.fastq -fastq_truncqual 2 -fastqout sample1_merged.fq

- **Quality filter**

$ usearch -fastq_filter sample1_merged.fastq -fastaout sample1_filtered.fa -fastq_maxee 1.0

- **Add the barcode**

$ sed "-es/^>\(.*\)/>\1;barcodelabel=Sample1;/" < sample1_filtered.fa > Sample1.fa

- **Add barcode name in front (optional)**

$ sed '-s/^>/>name_/' < input file name > output file name

- **Pool all samples together**

$ cat sample1.fa sample2.fa > reads.fa

Dereplication:

$ usearch -derep_fulllength reads.fa -output derep.fa -sizeout

Abundance sort and discard singletons:

$ usearch -sortbysize derep.fa -output sorted.fa -minsize 2

OTU clustering:

$ usearch -cluster_otus sorted.fa -otus otus1.fa

Chimera filtering using reference database:

$ usearch -uchime_ref otus1.fa -db rdp_gold.fa -strand plus -nonchimeras otus2.fa

Label OTU sequences OTU_1, OTU_2...

$ python ~/py/fasta_number.py otus2.fa OTU_ > otus.fa

Map reads (including singletons) back to OTU

$usearch -usearch_global reads.fa -db otus.fa -strand plus -id 0.97 -uc map.uc

Create OTU table:

$ python ~/py/uc2otutab.py map.uc > otu_table.txt

Create OUT table if barcode added in front (optional):

$ python ~/py/uc2otutab_mod.py map.uc > otu_table.txt

biom convert:

$ biom convert --table-type="OTU table" -i otu-table.txt -o otu-table.biom

(Importing UPARSE into QIIME pipeline for downstream analysis)

For both Fasta & Fastq format data using Oligotyping & MED:

(After alignment from QIIME or Uparse)

Clear those gaps:

$ o-trim-uninformative-columns-from-alignment A_aligned.fa

Entropy-analysis:

$ entropy-analysis A_aligned.fa-TRIMMED

Oligotyping (quick test):

$ oligotype A_aligned.fa-TRIMMED A_aligned.fa-TRIMMED-ENTROPY -c 10 --quick

Oligotyping (quick look before start):

$ o-stackbar.R mock-env-aligned-c2-s1-a1.0-A0-M0/ENVIRONMENT.txt -o mock --title Mock

Oligotyping:

$ oligotype A_aligned.fasta-TRIMMED A_aligned.fasta-TRIMMED-ENTROPY -c 12 -N 4 -a 3 --gen-html –m 454allsample_Sample.txt –o $PWD Oligotyping-c12-a03_sample

$ oligotype A_aligned.fasta-TRIMMED A_aligned.fasta-TRIMMED-ENTROPY -c 12 -N 4 -a 3 --gen-html –m 454allsample_Species.txt –o $PWD Oligotyping-c12-a03_species

$ oligotype A_aligned.fasta-TRIMMED A_aligned.fasta-TRIMMED-ENTROPY -c 12 -N 4 -a 3 --gen-html –m 454allsample_Region.txt –o $PWD Oligotyping-c12-a03_region

$ oligotype A_aligned.fasta-TRIMMED A_aligned.fasta-TRIMMED-ENTROPY -c 12 -N 4 -a 3 --gen-html –m 454allsample_Description.txt –o $PWD Oligotyping-c12-a03_SGMtest

MED:

$ decompose denoised_seqs_all_sample_aligned.fasta-TRIMMED -E 454allsample_sample.txt --gen-html -N 4 -m 1.2 -o $PWD/Med-m1.2-A0-M0-d4_sample

$ decompose denoised_seqs_all_sample_aligned.fasta-TRIMMED -E 454allsample_species.txt --gen-html -N 4 -m 1.2 -o $PWD/Med-m1.2-A0-M0-d4_species

$ decompose denoised_seqs_all_sample_aligned.fasta-TRIMMED -E 454allsample_region.txt --gen-html -N 4 -m 1.2 -o $PWD/Med-m1.2-A0-M0-d4_region

$ decompose denoised_seqs_all_sample_aligned.fasta-TRIMMED -E 454allsample_SGMtest.txt --gen-html -N 4 -m 1.2 -o $PWD/Med-m1.2-A0-M0-d4_SGMtest